Colloque international et école d'été
Albi, 10-14 juillet 2006
Organisé dans le cadre des Colloques d'Albi Langages et Signification (CALS)

Corpus en Lettres et Sciences sociales

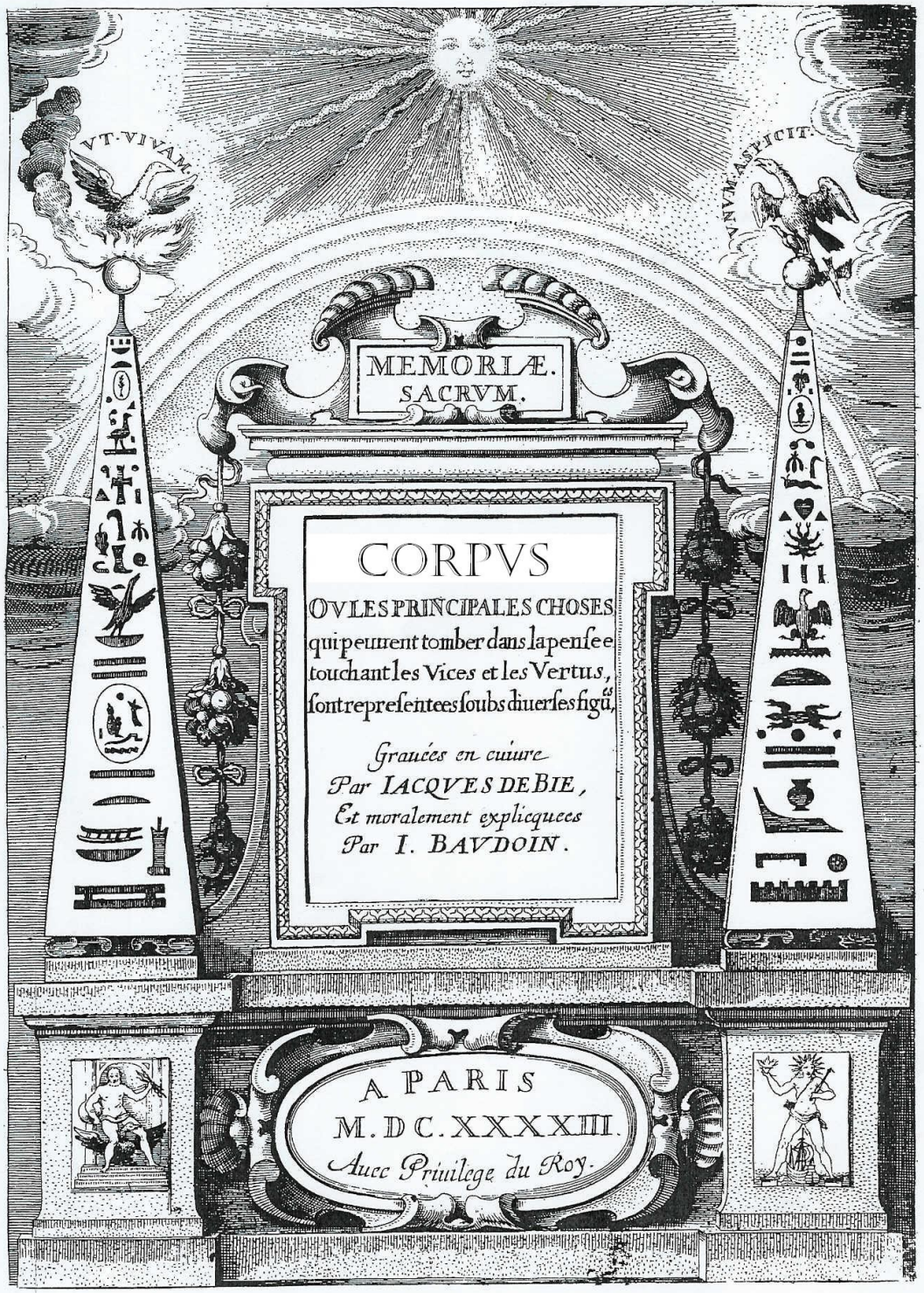
– Des documents numériques à l'interprétation

Actes

Publiés par Carine Duteil-Mougel et Baptiste Foulquié

2006

Editeur : Texto!
ISSN 1773-0120
Web : www.revue-texto.net


VT·VIVAM·
VNVM·ASPICIT·
MEMORIÆ.
SACRVM.
CORPVS
OV LES PRINCIPALES CHOSES
qui peuuent tomber dans la penſee
touchant les Vices et les Vertus,
ſont repreſentees ſoubs diuerſes figū,
Graucés en cuiure
Par IACQVES DE BIE,
Et moralement explicquees
Par I. BAVDOIN.
A PARIS
M. D C. XXXXIII.
Auec Priuilege du Roy.

TABLES DES MATIÈRES

Matthieu Perez
« Analyser la presse ancienne avec le progiciel *PHPRESS* : Le traitement numérique des faits divers de *L'Éclaireur de Nice*, 1928-1929 »

Augusta Mela et Mathieu Roche
« Des gloses de mot aux types de textes : un bilan différencié »

Lofti Abouda et Olivier Baude
« Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO »

Geoffrey Williams
« La linguistique et le corpus : une affaire prépositionnelle »

Huguette Rigot
« (En)-jeux de corpus pour la recherche en SHS. Enoncés, textes et documents »

Hassan Atifi, Christophe Lejeune, Goritsa Ninova, Manuel Zacklad
« Méthodologie transdisciplinaire de gestion du corpus pour les disciplines de l'interaction : recherche de principes directeurs »

Aurélien Bénel
« Porphyry au pays des paestans : usages d'un outil d'analyse qualitative de documents par des étudiantes de maîtrise en iconographie grecque »

Christophe Rey et Corinne Zaoui
« La résurrection du dictionnaire ancien par la déconstruction positive de l'informatique »

Driss Ablali
« Écrire en critique : exploration morpho-syntaxique sur corpus »

Magali Bigey
« Corpus et diachronie : de la constitution au traitement »

Carine Duteil-Mougel
« Groupement de textes et corpus : point de vue de linguiste »

David Cocksey
« L'œuvre complète numérique de Barbey d'Aurevilly »

Natalia Belozerova
« L'emploi des méthodes de la linguistique de corpus dans l'attribution des textes : les caractéristiques conceptuelles, sémantiques, épistémologiques du lexème « faith » dans les textes de Shakespeare »

Margareta Kastberg Sjöblom
« Étude quantitative des changements esthétiques et des variations génériques chez trois grands écrivains : analyse lexicométrique d'un corpus littéraire »

Jocelyne Le Ber
« L'adaptation comme contraction. Une analyse informatique d'Antigone »

Françoise Leriche
« Quel balisage pour les corpus numériques épistolaires ? De l'annotation traditionnelle du "document" à une analyse générique et pragmaticienne. »

Baptiste Foulquié
« De l'importance d'une théorie sémantique comme média entre corpus et analyse »

Simona Constantinovici
« L'œuvre poétique de Tudor Arghezi. La diversité du lexique et le problème du style »

Pierre Sadoulet
« Un instrument de lecture analytique. Présentation de *Corputex* »

Béatrice Akissi Boutin
« PFC-Abidjan : extension spatiale et disciplinaire d'un corpus »

Michel Ballabriga
Essai de synthèse

Liste des auteurs

*ABLALI Driss (Université de Franche-Comté)
ablali@u-paris10.fr

*ABOUDA Lofti & BAUDE Olivier (CORAL, Université d'Orléans)
labouda@univ-orleans.fr ; baude@wanadoo.fr

*ANDRÉ Virginie (CRAPEL-ATILF, Université Nancy 2)
Virginie.Andre@univ-nancy2.fr

*ARQUES Philippe (Professeur honoraire des universités)
philippe.arques@free.fr

*ATIFI Hassan, LEJEUNE Christophe, NINOVA Goritsa, & ZACKLAD Manuel
(TechCICO, Institut des Sciences et Technologies de l'Information, Troyes)
hassan.atifi@utt.fr ; christophe.lejeune@utt.fr ; goritsa.ninova@utt.fr ; manuel.zacklad@utt.fr

*BAUDE Olivier (CORAL, Université d'Orléans)
baude@wanadoo.fr

*BAUDOUIN Jean-Michel & PITA Juan
(Faculté de psychologie et des sciences de l'éducation, Université de Genève)
Jean-Michel.Baudouin@pse.unige.ch

*BELOVA Svetlana (Université d'Etat de Tioumen, Russie)
s_belok@hotmail.com

*BELOZEROVA Natalia (Université d'Etat de Tioumen, Russie)
nbelozerova@utmn.ru

*BENEL Aurélien (Tech-CICO, Université de Technologie de Troyes)
aurelien.benel@utt.fr

*BIGEY Magali (Université de Franche-Comté, LASELDI)
magali.bigey@univ-fcomte.fr

*BOUTIN Béatrice Akissi (ERSS, Université Toulouse 2)
boubeaki@hotmail.com

*BRUNET Étienne (Université de Nice, BCL)
brunet@unice.fr

*COCKSEY David (LLA, Université Toulouse 2)
david.cocksey@free.fr

*CONSTANTINOVICI Simona (Université de l'Ouest, Timişoara, Roumanie)
simonadiana@hotmail.com

*CORNUÉJOLS Martine (Chercheur associé MoDyCo, Université Paris X)
m.cornuejols@laposte.net

*DESQUINABO Nicolas & BECQUERET Nicolas (Paris 3 - Sorbonne Nouvelle)
nicodeski@yahoo.fr

*DUTEIL-MOUGEL Carine (ATILF, Nancy-Université, CNRS)
Carine.DUTEIL@wanadoo.fr

*FOULQUIÉ Baptiste (CPST, Université Toulouse 2)
btistou@hotmail.com

*HAJLAOUI Najeh (CLIPS, GETA, IMAG, Université Joseph Fourier Grenoble)
najeh.hajlaoui@imag.fr

*KASTBERG SJÖBLOM Margareta (ILF-CNRS, BCL, Université de Nice)
kastberg@wanadoo.fr

*KREBS Constance (Paris 3, Censier-Sorbonne Nouvelle)
constance.krebs@noos.fr

*KUTUZOV Andrey (Université d'Etat de Tioumen, Russie)
tyumenkender@gmail.com

*LE BER Jocelyne (Collège militaire royal du Canada)
jocelyne.le.ber@rmc.ca

*LERICHE Françoise (Grenoble 3 et I.T.E.M.)
francoise.leriche@wanadoo.fr

*LOISEAU Sylvain (MoDyCo, Université Paris X)
sylvain.loiseau@wanadoo.fr

*LORTAL Gaëlle, TODIRASCU-COURTIER Amalia & LEWKOWICZ Myriam
(Tech-CICO, Université de Technologie de Troyes) et (LILPA, Université Strasbourg)
lortal@utt.fr ; amalia.todirascu@umb.u-strasbg.fr ; lewkowicz@utt.fr

*MANGE Jessica (IUT GEA), MARCHAND Pascal (LERASS / IUT Information & communication) &
SALEM André (EA2290 SYLED – CLA2T - Paris 3)
jessica.mange@gmail.com ; pascal.marchand@iut-tlse3.fr ; salem@msh-paris.fr

*MAYAFFRE Damon (BCL, CNRS, Université de Nice)
damonmayaffre@wanadoo.fr

*MELA Augusta & ROCHE Mathieu (LIRMM, Univ. de Montpellier)
Augusta.Mela@univ-montp3.fr ; mathieu.roche@lirmm.fr

*NABTI Karima (Université d'Alger)
karima_nabti@hotmail.com

*NTABONA Adrien (Université du Burundi)
ntabona_a@yahoo.fr

*PEREZ Matthieu (CMMC, Université de Nice - Sophia Antipolis)
matt.perez@wanadoo.fr

*PESCHEUX Marion (Université Jean Monnet, St Etienne, CERCI, Nantes)
marionpescheux@free.fr

*PINCEMIN Bénédicte (Laboratoire de Linguistique Informatique de Paris 13, CNRS)
benie@club-internet.fr

*PLISSONNEAU Gersende (IUFM, Grenoble)
gersende.plissonneau@wanadoo.fr

*POUDAT Céline (CORAL, Université d'Orléans)
celine.poudat@univ-orleans.fr

*PRIGNITZ Gisèle (ERSS, Université de Pau et des Pays de l'Adour)
gisele.prignitz@univ-pau.fr

*RASTIER François (CNRS, Paris)
Lpe2@ext.jussieu.fr

*REY Christophe & ZAOUI Corinne (DELIC, Université de Provence)
christophe.rey@up.univ-aix.fr zaoui@up.univ-aix.fr

*RIGOT Huguette (Paris X, INRP)
huguette.rigot@paris7.jussieu.fr

*SADOULET Pierre (CIEREC, Université de Saint-Étienne)
pierre.sadoulet@club-internet.fr

*TAURISSON Dominique (SHADYC, EHESS-CNRS)
dominique.taurisson@univmed.fr

*VALETTE Mathieu (ATILF, Nancy-Université, CNRS)
mathieu.valette@atilf.fr

*WILLIAMS Geoffrey (Université de Bretagne Sud, Lorient)
geoffrey.williams@wanadoo.fr

*WILLIAMS John (lexicographe / enseignant indépendant)
johnwhoever@wanadoo.fr

*YOUSFI Abdellah & EL-JIHAD Abdelhamid (IERA, Université Mohammed V Souissi Rabat-Maroc)
yousfi240ma@yahoo.fr ; eljihad@ifrance.com

AVANT PROPOS

François RASTIER
UMR 7114 et ERTIM (Inalco)

Un beau jour du printemps 2004, une journaliste du CNRS vint me trouver et me demanda de lui parler de l'amour au XXI^e^ siècle. Sur ce sujet éminemment consensuel, le *Journal du CNRS* préparait un dossier interdisciplinaire et l'ouvrage que j'avais dirigé quelques années auparavant, *L'analyse des données textuelles — L'exemple des sentiments dans le roman français (1820-1970)* avait semblé me qualifier pour traiter de cette question.
Conscient de mes obligations statutaires, je m'efforçai de répondre, mais je le fis par la question : « Dans quel corpus ? ». Devant le désarroi qui se peignit sur le visage avenant de mon interlocutrice, je me lançai dans des justifications : pour nous, malheureux linguistes, l'amour n'existait que dans les textes et variait avec les discours, les genres et les auteurs. Ainsi n'avait-il rien de commun dans le roman du XIX^e^ siècle, où *amour* trouve pour antonymes *argent* et *mariage*, et dans la poésie de la même époque, où l'argent et le mariage restent évidemment absents. Faute d'avoir eu la présence d'esprit de constituer un corpus sur l'amour en ce siècle naissant, je dus enfin confesser mon incompétence. Tout cela dut paraître bien décevant et il n'en résulta qu'un maigre entrefilet dont je suis confus de n'avoir gardé aucun souvenir.
Il me parut donc nécessaire d'entreprendre une action de communication, non plus à propos de l'amour, sujet apparemment porteur, mais des corpus en lettres et en sciences humaines. Je proposai à Michel Ballabriga et Pierre Marillaud d'organiser à ce propos un colloque, dont voici à grands traits l'argument.

De nombreuses collectivités sont de longue date engagées dans une réflexion sur la numérisation et l'analyse assistée des documents : outre bien entendu les sciences de l'information, il faut mentionner entre autres l'histoire, la sociologie, la linguistique, l'archéologie, les études littéraires.
La constitution et l'analyse de corpus est en passe de modifier les pratiques voire les théories en lettres et sciences sociales. Toutes les disciplines ont maintenant affaire à des documents numériques, et cela engage pour elles un nouveau rapport à l'empirique. En outre, la numérisation des textes scientifiques eux-mêmes permet un retour réflexif sur leur élaboration et leurs parcours d'interprétation. Les nouveaux modes d'accès aux documents engagent-ils de nouvelles formes d'élaboration des connaissances ?
Les nouvelles initiatives prises au plan national et international peuvent devenir l'occasion et donner les moyens d'un projet fédérateur pour les lettres et les sciences sociales.
Aussi ce colloque ouvert entend-il renforcer des liens et favoriser de nouvelles rencontres d'enseignants et de chercheurs de ces disciplines avec ceux des collectivités de la linguistique de corpus et du document numérique. Sans trop d'égard pour l'objectivisme ordinaire, il traite des problèmes philologiques et herméneutiques que pose le travail sur des corpus numériques en fonction des tâches et des disciplines. Il s'attache par exemple à la typologie des genres et discours, à la description de formes et de fonds sémantiques, au repérage de thèmes, à la caractérisation et à l'évolution de concepts, à l'étude des corrélations contenu/expression.
Au plan pratique, il aborde les questions que posent le recueil, l'établissement, le codage, l'étiquetage, le traitement des corpus et leur édition électronique.

On connaît les travers ordinaires des colloques disciplinaires (vedettariat, meurtre du congénère) et des colloques interdisciplinaires (métadiscours grandiloquent) : pluridisciplinaire sans prétendre mettre en scène une interdisciplinarité sans rivages, celui-ci s'est tenu dans une atmosphère sereine de doute enthousiaste, chacun ayant le souci de présenter sa problématique sans en cacher les limites ni négliger les difficultés liées à la constitution des corpus et à l'interprétation des résultats. Des démonstrations de logiciels ont été assurées ainsi que des initiations aux problématiques propres des différentes disciplines concernées.
Le doute positif relève de l'attitude critique nécessaire à toute problématisation scientifique. Il reçoit ici un contenu nouveau, car avec les corpus numériques, les sciences de la culture trouvent de nouvelles perspectives épistémologiques et méthodologiques, alors qu'elles se trouvent affrontées à des programmes réductionnistes de naturalisation des cultures.

L'objection classique formulée contre leur scientificité tient au caractère non répétable des événements : comme en sociologie, en ethnologie, en psychologie sociale voire en linguistique de l'oral, la présence même de l'enquêteur modifie la situation, on conclut que les sciences de la culture n'auraient donc pas la possibilité d'identifier des causes déterminantes et donc des lois. Or selon le préjugé scientiste qui sous-tend les programmes de naturalisation, la condition nécessaire de la scientificité reste la formulation de lois causales – qu'il faudrait alors chercher dans les substrats physiologiques, neuronaux ou génétiques (cf. Sperber et « l'épidémiologie des représentations » comme explication globale de la culture).
À la classique dualité induction/déduction des disciplines d'observation, le renouvellement méthodologique favorisé par les corpus numériques engage à substituer le cycle suivant : (i) recueil d'information et production des données ; (ii) élaboration de documents scientifiques ; (iii) traitement instrumenté des corpus ; (iv) interprétation des résultats.
La puissance propre de ce dispositif permet de faire émerger de nouveaux observables inaccessibles autrement : par exemple, la phonostylistique, jadis condamnée à l'intuition, se voit à présent pourvue de moyens d'investigation par les statistiques sur corpus phonétisés. En outre, l'utilisation d'une instrumentation scientifique (analyseurs, étiqueteurs, etc.) participe du processus d'objectivation : les objets culturels ont beau dépendre de leur conditions d'élaboration et d'interprétation, les valeurs qu'ils concrétisent peuvent cependant être objectivées comme des faits.
La linguistique de corpus pourvoit ainsi la linguistique d'un domaine où élaborer des instruments et définir une méthode expérimentale propre : elle ouvre aussi des champs d'application nouveaux et engage un nouveau mode d'articulation entre théorie et pratique. D'une part, alors que la linguistique théorique – sans corpus – portait, en extrapolant quelques observations sur des exemples souvent forgés, des jugements universels sur le langage, la linguistique de corpus, sans renoncer à l'élaboration théorique, en limite la portée aux corpus étudiés, et, sans se satisfaire de la seule démarche déductive, procède par essais et erreurs.
En 1999, Chomsky, auteur d'une grammaire universelle, déclarait que la linguistique de corpus n'existait pas, alors même qu'elle était déjà en plein essor : il signalait par ce petit meurtre symbolique qu'elle restait inconcevable pour la linguistique de fauteuil et qu'une rupture épistémologique était en cours. Elle jouit d'une portée générale : en bref, la recherche part d'une diversité constatée, l'unifie dans le point de vue qui préside à la collection du corpus, éprouve son objectivité par l'investigation instrumentée. L'unité, ou du moins la régularité, sera créditée au système, la diversité irréductible au corpus. Ainsi l'opposition entre l'unité substantielle et l'irrégularité accidentelle peut-elle être dépassée dans la description des normes, dont les plus générales, parmi l'ensemble des corpus étudiés, seront considérées comme propres à la langue.

Sans prétendre tirer un bilan prématuré, il semble que cette situation nouvelle conduit à une reconception de la dualité entre linguistique de la langue et linguistique de la parole, qu'il est de tradition d'opposer, tant chez Bally que chez Benveniste, tant en linguistique de l'énonciation qu'en pragmatique, alors que chez Saussure elles sont parfaitement complémentaires.
On a trop souvent réduit les langues à des dictionnaires et des grammaires, voire à des syntaxes. Il faut cependant tenir compte, outre du *système*, du *corpu*s (corpus de travail et corpus de référence), de l'*archive* (de la langue historique), enfin des *pratiques* sociales où s'effectuent les activités linguistiques. Pour l'essentiel, une langue repose sur la dualité entre un *système* (condition nécessaire mais non suffisante pour produire et interpréter des textes) et un *corpus* de textes écrits ou oraux[1].
La dualité entre corpus et système n'a rien d'une contradiction : elle est prise dans la dynamique qui constitue la langue dans son histoire. Aussi ne saurait-on assimiler la langue historique à la langue fonctionnelle (celle qui fonctionne ici et maintenant) en négligeant que la langue historique détermine la langue fonctionnelle dans ses structures et ses contenus. Le corpus sert de

[1] Dans le corpus d'une langue, les *œuvres* tiennent une place particulière parce qu'elles sont valorisées et prennent le rang de parangons : par exemple l'italien n'est pas moins la langue de Dante que Dante le parangon historique qui a présidé à la formation de la langue italienne en tant que langue de culture (supplantant l'occitan). Plus généralement, bien des expressions, dictons et proverbes renvoient aux poètes, législateurs et historiens d'autrefois : ainsi, en chinois, des expressions en quatre caractères qui fourmillent à l'écrit comme à l'oral.

médiation entre la langue historique et la langue fonctionnelle, et les textes qui n'appartiennent plus qu'à la langue historique entrent dans l'archive.
En parlant de corpus et non de signes, nous soulignons que la langue n'est pas un système de signes comme le serait un code ; Saussure, à qui l'on prête cette définition, ne l'a jamais formulée. Un signe au demeurant n'a pas de définition intrinsèque : il n'est qu'un passage, certes réduit, d'un ou plusieurs textes auxquels il renvoie. Bref, une langue est faite d'un corpus de textes et d'un système : le système reconstitué par les linguistes a le statut d'une hypothèse rationnelle formulée à partir des régularités observées dans le corpus. Entre le corpus et le système, les normes assurent un rôle de médiation : ancrées dans les pratiques sociales, les normes de discours, de genre et de style témoignent de l'incidence des pratiques sociales sur les textes qui en relèvent[1].

L'essor de la linguistique de corpus conduit à préciser le rapport entre textes et documents. Alors que la grammaire travaillait sur l'écrit (son nom même l'indique, littéralement), l'oral est une conquête récente de la linguistique ; encore faut-il qu'il soit fixé sur un support, par enregistrement ou transcription, pour devenir l'objet des débats et conjectures propres à l'investigation scientifique. Textes oraux et écrits trouvent leur première unité dans leur statut de documents.
Plus généralement, les différences entre texte et document, bibliothèque et archive, linguistique de corpus et philologie numérique, sont en train de devenir relatives. Le support numérique ne garantit aucune identité à soi : la restitution de l'inscription est sensible aux formats, aux logiciels de visualisation dont les standards évoluent, si bien que la notion philologique d'herméneutique matérielle doit ici être dépouillée de tout attendu substantiel.
En perdant son unicité, le document numérique se dépouille des qualités du document unique de l'archiviste : authentifiable, doué par sa continuité matérielle d'une intégrité (même quand il est fragmentaire), non reproductible, faisant autorité. L'affichage par pixel détruit toute continuité matérielle qui empêchait les falsifications. Alors qu'une critique initiale suffisait à établir le document, il faut à présent une critique indéfinie pour maintenir une fiabilité. L'établissement des significations doit souvent passer par une succession de versions, dont chacune est le support et le résultat d'une opération de lecture. Changeant de régime, l'objectivation doit être indéfiniment progressive sans pouvoir jamais être considérée comme établie, ce qui engage à rompre avec l'objectivisme pour promouvoir une objectivation critique indéfinie.
Toutefois, ce que le document perd en stabilité, il le gagne en biais d'interrogation. Les logiciels imposent une réflexion théorique sur l'étiquetage, sur les rapports entre méthodes qualitatives et quantitatives : on peut par exemple croiser les résultats de plusieurs méthodes pour faire apparaître de nouveaux observables. C'est autant aux « gens du texte » qu'aux informaticiens de faire des propositions sur ce point : pour aborder ces questions, la voie technologique et la voie épistémologique n'ont rien de contradictoire.
C'est par la méthodologie comparative que l'on va pouvoir exploiter les possibilités techniques actuelles. Pour fonder cette méthode, lui permettre d'évoluer et lui fixer des objectifs de connaissance, il faut aussi que la linguistique assume sa place parmi les sciences de la culture.
En renouant avec les corpus, la linguistique renoue nécessairement avec les textes, donc avec la philologie et avec l'herméneutique : la philologie pour les établir et les documenter, l'herméneutique pour les interpréter, y compris dans leur dimension intertextuelle.

Si nous avons beaucoup appris pendant ce colloque, nous le devons aussi au CALS et à ses animateurs : j'ai plaisir au nom de tous les participants à remercier Béatrix et Pierre Marillaud pour leur accueil chaleureux et leur organisation sans faille qui nous ont permis de vivre un moment d'utopie bien présente.

[1] Un texte en effet ne peut pas être produit par un système, comme l'a montré l'échec de la grammaire générative appliquée à des systèmes de génération automatique.

PHILOLOGIE ET/OU HERMÉNEUTIQUE NUMÉRIQUE : NOUVEAUX CONCEPTS POUR DE NOUVELLES PRATIQUES ?

Damon MAYAFFRE
CNRS / UMR 6039, Bases, Corpus et Langage (Nice)

SOMMAIRE

Résumé *: L'enjeu des sciences du texte est moins d'administrer la preuve que de contrôler l'interprétation. Hors de l'obscurantisme théologique, il faut admettre en effet que les textes, et les corpus qui en informent le sens, n'ont point de Vérité mais de multiples compréhensions. Selon une pensée attribuée à Foucault, la vérité d'un texte est d'abord et seulement ce qu'on dit de lui, et déjà Chladenius remarquait que, loin de la stricte intentionnalité des auteurs, « l'on peut, lorsqu'on cherche à comprendre leurs écrits, former des pensées qui n'étaient pas venues à l'esprit de l'auteur » [Chladenius cité par Szondi 1989 : 32].*
Seulement, sauf à verser dans un subjectivisme débridé et une interprétation divinatoire, « ce qu'on dit des textes » et ces « pensées » qu'il est permis d'avoir à propos d'eux, doivent être étayés, vérifiables, contrôlés. Cela passe par une composition/organisation ad hoc des corpus, une prise en considération minutieuse du matériel linguistique qui les constitue, et une démarche heuristique rigoureuse. Dans les trois cas, la révolution numérique apporte des réponses adéquates.

Note liminaire

Adoptons le parti pris dans cette contribution d'agglutiner philologie et herméneutique. Leur définition/spécification mériterait un article à part entière. Leur association –notée ici sous la forme relâchée et consensuelle : « philologie et/ou herméneutique »– témoigne simplement que nous ne considérons pas seulement la philologie, de manière réductrice, comme une technique d'établissement des textes, mais aussi, pour ce faire, comme l'art de leur appréhension, c'est-à-dire de leur compréhension ; c'est-à-dire de leur interprétation. De la même manière, on ne désignera pas uniquement par herméneutique, l'interprétation théologique, philosophique, allégorique, etc. des textes, mais l'art d'en établir non seulement le sens profond mais l'origine exacte, le fond supposé mais la forme attestée, le contenu mais l'expression ; l'esprit des textes donc, mais avant cela, nécessairement, la lettre.

En un mot, prises chacune dans une acception pleine, philologie et herméneutique sont indispensables l'une à l'autre ; partie prenante l'une de l'autre. Longtemps artificiellement séparées, pour des raisons historico-épistémologiques plurielles que certains auteurs ont décrites, elles peuvent se réconcilier à la faveur de la révolution numérique dont il sera question dans cette contribution : il s'agirait même d'une des conséquences les plus heureuses, au sein des sciences de la culture, de la révolution numérique en question.

Cette position liminaire nous est directement inspirée par la lecture de P. Szondi (avant propos de J. Bollack), Introduction à l'Herméneutique Littéraire. De Chladenius à Schleirmacher (Cerf, trad. 1989) où l'idée d'une herméneutique « critique » ou « matérielle » –à défaut, directement, d'une herméneutique philologique– est défendue. Et par celle de F. Rastier, Arts et Sciences du texte (Puf, 2001) qui semble être le principal penseur contemporain à établir le « projet d'unifier l'herméneutique et la philologie » (p. 276 ; cf. aussi p. 2) quand bien même ce projet passerait par une reconsidération de l'obje(c)t(if) de la linguistique et une prise en considération novatrice des possibles du numérique en matière de textes, de corpus, de procédures heuristiques, d'outils de recherche, de formalisation des parcours interprétatifs.

Introduction

Ce propos débute sur un constat empirique, d'ordre personnel, mais qui est, semble-t-il, suffisamment partagé aujourd'hui en SHS pour être généralisé.

Dans le cadre d'une étude linguistico-historique du langage politique français, j'ai étudié, au milieu des années 1990, des corpus textuels *papiers* –puisés par exemple dans l'œuvre de Maurice Thorez, éditée en plusieurs volumes par les Editions sociales. Je poursuis aujourd'hui mon travail par l'étude de corpus textuels *numériques* –puisés par exemple dans l'œuvre de Jacques Chirac éditée en plusieurs millions d'octets par le site officiel de l'Elysée.

Au terme de cette évolution, il apparaît que ce qui pouvait être considéré comme un simple changement du support de l'objet de recherche (des corpus textuels donc, ici composés d'une collection de textes politiques contemporains) entraîne un changement de la perception de sa nature, de la nature de ses composants (les textes) et, par là, un changement de leur compréhension-interprétation.

Pour cette raison, il faut, sans crainte d'apparaître naïvement moderne, affirmer, en France, avec [Rastier, 2001] dès le début du siècle, plus modestement avec [Mayaffre 2002-a], récemment avec [Viprey, 2005] et encore, cette année, avec [Adam, 2006] que la philologie et/ou herméneutique numériques révolutionnent non seulement notre rapport aux textes et à la textualité, mais aussi nos pratiques heuristiques quotidiennes, mais encore, tout simplement, nos connaissances et notre appréhension de la culture (textuelle) humaine.

La question est aujourd'hui moins de savoir si la révolution numérique est aussi importante que celle de l'imprimerie dont on sait le rôle dans la propagation de l'humanisme, de la Réforme et des Lumières –d'évidence elle l'est ; aussi importante et plus rapide– que de savoir si une révolution, fût-elle scientifique ou culturelle, peut, au-delà de se vivre, se théoriser ?

La question est surtout de savoir si, comme toutes les révolutions, la révolution du tout numérique –ici des corpus textuels numériques– saura résister au double danger qui la menace sur sa gauche et sur sa droite par la surenchère ou la restauration.

À sa gauche, le passage du papier à l'électronique a entraîné le développement de l'Analyse de Données Textuelles (ADT) et, de manière plus désincarnée, du Traitement Automatique des Langues (TAL). Or ces pratiques, si elles ne devaient être que techniques ou algorithmiques, et devaient toujours surenchérir vers l'automatisme, souffriraient d'un déficit philologique pour la première, et d'un déni philologique pour la seconde. Il y aurait là, autour des textes, un divorce désastreux entre elles et les humanités.

À sa droite, les tenants de l'ancien régime papier continuent une longue tradition qui n'a aucune raison de s'éteindre. En dépit d'une évolution que l'on peut juger comme inéluctable, la lecture empathique ou intuitive des textes –lecture pré-saussurienne d'une part qui fait fi des apports des sciences du langage, lecture anté-numérique d'autre part qui ignore les possibilités des nouveaux supports, des nouveaux médias, des nouveaux outils–, demeure encore aujourd'hui majoritaire en SHS et en appelle seulement, comme suprême argument, à la sensibilité et l'érudition de l'analyste. Les logiciels d'analyse de données textuelles par exemple restent au mieux des gadgets d'appoint dans l'art d'interpréter les textes ; au pire totalement ignorés. Le divorce serait alors à la fois social et scientifique : aux internautes d'un côté et aux linguistes spécialisés de l'autre le loisir de manipuler, télécharger, formaliser et disséquer les textes, aux lettrés érudits le privilège supposé de les goûter et de les comprendre.

1. Visions sur les corpus textuels numériques

Les textes sont des artefacts, les corpus des construits. Ces deux postulats de la linguistique textuelle et de la linguistique de corpus, difficilement contestables, et aux conséquences épistémologiques multiples, ne sont pas strictement liés à la révolution numérique. Mais il n'est pas un hasard si [Viprey 2005] les rappelle à l'occasion de son article *Philologie numérique et herméneutique intégrative* dans lequel il décrit les apports décisifs du support digital dans les sciences et arts du texte.

Tout se passe en effet comme si la transition vers le numérique avait rendu incontournables et impérieuses quelques évidences philologiques et/ou herméneutiques oubliées.

1.1. Le texte est un artefact

Le texte est un artefact (*artis factum* : fait de l'art), *phénomène d'origine humaine, artificielle*, comme l'indique la définition. La passage du papier au numérique, le travail technique et quotidien

de saisie par exemple[1], la simple lecture du texte sur *son* écran *via* l'ascenseur de *son* traitement de texte[2], sans parler de la réflexion théorique et pratique sur l'édition numérique, les options de codage, de balisage, d'étiquetage, tout cela nous fait rompre avec l'idée qu'il existerait un texte naturel, dont la forme intangible serait le folio ou le livre avec sa couverture et sa pagination. Bien sûr, la philologie traditionnelle, en insistant sur les différentes éditions et en développant le comparatisme non hiérarchisé [voir récemment Heidmann 2005 ; Adam 2005], avait prévenu contre la naturalisation abusive d'un texte source et réifié. Mais la philologie numérique expérimente cette réalité tous les jours en relativisant la forme textuelle.
Cette relativisation peut aller loin dans l'Analyse de données textuelles puisque les logiciels permettent de faire apparaître, à l'écran, le texte sous différentes formes conventionnelles. La convention la mieux établie est la surface graphique ; et la stabilité relative de l'apparence graphique ne devra pas nous faire perdre de vue qu'il ne s'agit là que d'une convention. Mais à côté du texte graphique, nu ou brut, le texte lemmatisé et étiqueté peut aussi se laisser voir à l'écran. *HYPERBASE*, articulé au lemmatiseur *CORDIAL*, permet ainsi de juxtaposer, dans un même mouvement, plusieurs conventions [*illustration 1 : Texte brut et texte lemmatisé de Jacques Chirac (14 juillet 1995, conférence de presse)*. Dans la fenêtre de gauche le texte brut ; dans la fenêtre de droite le texte lemmatisé où tous les mots graphiques ont été ramenés à leur lemme d'origine et où chaque lemme est suivi d'un code de 0 à 9 pour les grandes catégories grammaticales (1 = verbe, 2 = substantif, etc.)].

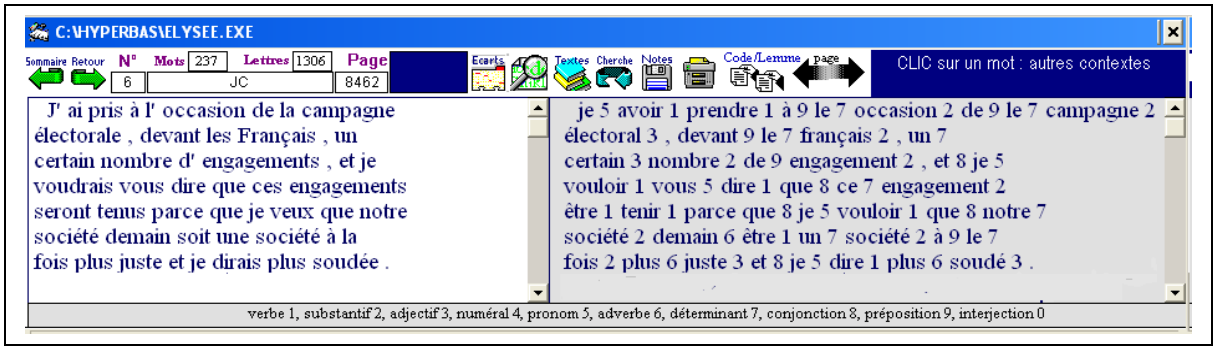


Illustration 1 : Texte brut et texte lemmatisé d'un discours de J. Chirac (14 juillet 1995)

Ce que nous voulons montrer, par cet exemple, c'est que le numérique en multipliant les mises en forme des textes propose une autre vision du texte. Un texte anti-naturel donc, dématérialisé –virtuel pourrait-on dire commodément–, dont les contours physiques tels que perçus depuis des siècles sont abolis, et la structure et le contenu –entendons, pour faire simple : la textualité– reconsidérés.
Il faut insister, ici, sur l'aspect le plus novateur de ces visions alternatives du texte que peut entraîner le numérique : *le dépassement/complément de la linéarité*.
La plupart des définitions du texte insistent en effet sur l'unité dynamique qu'il représente. La plus significative, dans ce sens, est celle que donnent [Détrie, Siblot, Vérine, 2001] dont on souligne les éléments saillants :

> *Un texte est une **suite** d'énoncés oraux ou écrits posés par leur producteur –et destinés à être reconnus par leur(s) destinataire(s)– comme un ensemble cohérent **progressant vers***

[1] Que saisit-on exactement ? Le corps du texte seulement ? La couverture et les en-têtes ? Et quelle édition choisir ? Quel format de restitution demander au logiciel de reconnaissance de caractères ? Même lorsqu'ils sont tirés de documents papiers, les documents électroniques ne peuvent être la reproduction exacte d'originaux, à moins de seulement photographier les textes. Mais précisément, nous aurions alors affaire à des images et non plus à des textes. Les derniers développements du format PDF sont intéressants à ce sujet. Pendant longtemps le PDF était la reproduction fidèle et intangible du format papier. Seulement, la manipulation de ces fichiers images a très vite parue rigide pour l'utilisateur. Aussi, il est désormais possible de transformer avec *PDF Converter* l'image en texte,… et le caractère intangible du contenu se trouve remis en cause.
[2] La multiplicité et la personnalisation des écrans d'ordinateurs (taille, forme, résolution) et des traitements de texte (quelle police par défaut ? Mode page ou mode normal ?) font qu'aucun texte n'apparaît désormais au lecteur sous la même forme.

une fin *et parvenant à constituer une complétude de sens. [Détrie, Siblot, Vérine, 2001 : 349]*

« Suite » [cf. aussi Rastier 2001 : 21], « plan » [Adam 1999 : 5] : la linguistique textuelle insiste, non sans argument, sur la linéarité, le déroulement séquentiel, l'enchaînement, la progression, la *cohésion*[1] d'un texte.

Pourtant le support et l'outillage électroniques permettent à moindre coup de doubler le point de vue de la linéarité par d'autres points de vue que proposent d'autres types de lecture.

Ont été relevées, dans [Mayaffre 2002-a], trois lectures électroniques complémentaires à la lecture oculaire linéaire traditionnelle : lecture quantitative (complémentaire de la lecture qualitative), lecture paradigmatique (complémentaire de la lecture syntagmatique), lecture hypertextuelle (complémentaire de la lecture textuelle). Et si l'on insiste sur la dimension complémentaire de ces approches, c'est que l'opposition entre numérique et oculaire n'a pas lieu d'être : la philologie et/ou herméneutique numérique entend prolonger, mais aucunement abolir, l'analyse de texte habituelle. *Hyperbase* par exemple s'applique à croiser l'approche quantitative du texte et l'approche qualitative. Aux fonctions statistiques, caractéristiques du logiciel (« spécificités », « accroissement lexical », « distance intertextuelle », « corrélation chronologique », « richesse du vocabulaire », etc.), se combinent des fonctions d'exploration qualitative (« lecture », « concordance », « contexte »). Surtout, ces fonctions tentent de se féconder, de se juxtaposer, de se superposer dans l'ergonomie même du logiciel. Le bouton « Lecture », par exemple, donne accès au texte tel qu'il a été saisi et invite à une lecture linéaire, qualitative, intuitive, ordinaire en faisant défiler le texte, dans sa continuité, comme on tourne les pages d'un ouvrage. Pourtant si l'on actionne le bouton « Ecarts », alors le texte « naturel » s'anime et met en relief les mots qui sont caractéristiques statistiquement de la partie du corpus concernée. [*Illustration 2 : Lecture assistée d'un discours J. Chirac (3 avril 2002, interview télévisée)*. À gauche, le texte est lisse. À droite le texte est en relief avec les mots sur-utilisés par Chirac (par rapport à l'ensemble du corpus présidentiel 1958-2002) soulignés].

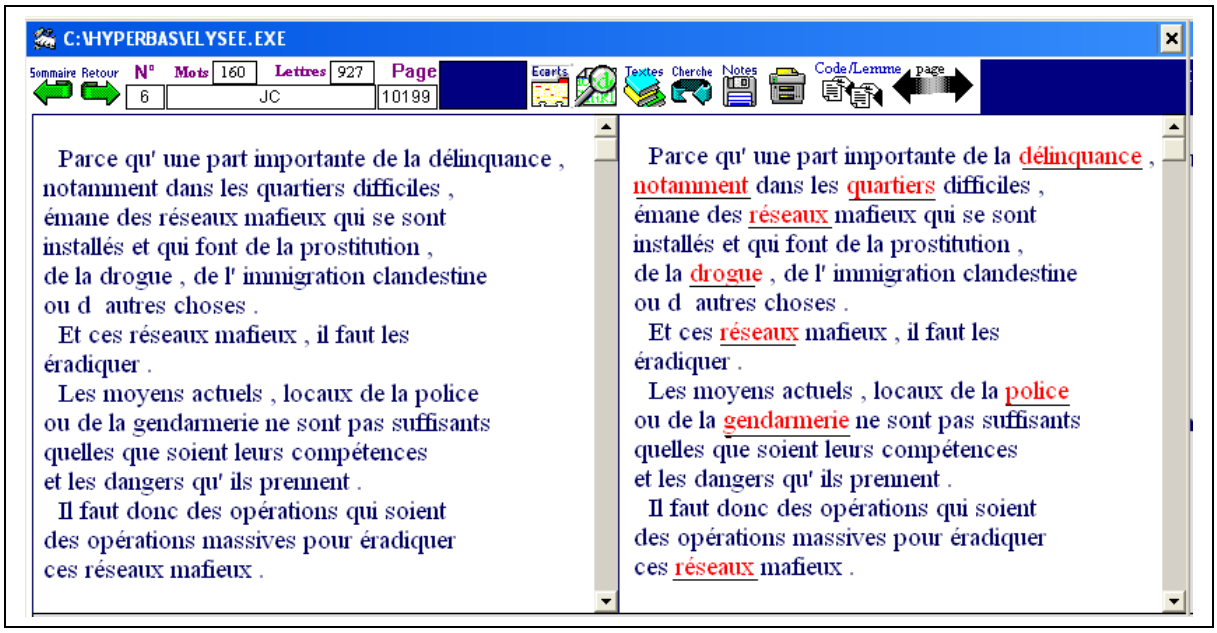


Illustration 2 : Lecture assistée d'un discours de J. Chirac (3 avril 2002, interview télévisée)

Le lecteur pourra donc lire et compter dans un seul élan. Sa lecture intuitive sera *assistée* par la statistique selon le mot d'Etienne Brunet, et l'esprit mis en alerte sur les mots quantitativement discriminants de telle ou telle partie du corpus. Loin d'être un gadget, la fonction « Ecarts » fond, en espérant les réconcilier, deux approches désormais bien établies du texte, deux traditions longtemps séparées : le scriptural et la métrique.

[1] À nous de montrer avec [Viprey 2005 : 66 et ss] que la *cohésion* d'un texte ne désigne pas seulement « sa **continuité** sémantique » [Détrie, Siblot, Vérine, 2001 : 57] ou sa « **progression** thématique » [Charaudeau et Maingueneau, 2002 : 99]. En attendant, le concept s'inscrit bien dans la vision linéaire du texte.

Lecture quantitative, lecture paradigmatique (par le biais d'index notamment), lecture hypertextuelle (par le jeu des liens et des renvois), disions-nous, en complément de la lecture linéaire usuelle : les mots étaient peut-être maladroits et [Viprey 2005] résume le changement en des termes plus percutants. Il fixe comme objectif à la philologie et/ou herméneutique numérique de combiner la lecture linéaire à des lectures *tabulaire et réticulaire*.
De fait, les logiciels d'Analyse de données textuelles, notamment ceux qui privilégient l'approche quantitative, commencent par faire exploser la linéarité du texte pour présenter leurs données en tableaux : tableaux alphabétiques, tableaux de fréquences, tableaux de distances, etc. Ces tableaux ne prétendent certes pas être le texte, mais ils sont une vision systématique et organisée –après l'explosion, le rangement– de la matière textuelle et deviennent les matrices sur lesquelles nos interprétations seront fondées.
Plus subtilement, l'enjeu le plus complexe de l'Analyse de données textuelles est de déceler les relations –relations autres que syntaxiques– que les items linguistiques entretiennent entre eux, non dans la phrase mais dans le texte en sa globalité. Texte, textualité, texture : l'objectif est de renouer avec l'étymologie même de ces mots et de démêler les trames et les entrelacs sous-jacents. Vision réticulaire donc des textes et des corpus qui met à jour les réseaux lexicaux pour (re)construire les thématiques, les isotopies ou isotropies récurrentes. De manière magistrale, [Viprey 2005], outillé par l'AFC, illustre le propos par l'étude de « l'organisation micro-distributionnelle » [*ibid* : 61] des vocables dans le Monde Diplomatique grâce à l'étude du « système de collocation » [*ibid* : 62]. Et la fonction « Thème » d'*HYPERBASE* appliquée au corpus présidentiel français (1958-2002), permet de repérer les mots attirés par un mot pôle et de reconstituer ainsi dans une approche micro d'un macro corpus (la fenêtre d'étude étant le simple paragraphe et le corpus embrassé comptant plus de 500 discours) le système des co-occurrences qui font *nombre* c'est-à-dire *sens* [*illustration 3 : Environnement lexical du mot « mondialisation » dans le discours de J. Chirac*. Le tableau fait apparaître par ordre hiérarchique les mots qui sont le plus attirés par « mondialisation ». Trois traits isotopiques du discours peuvent ainsi être distingués. Dans un propos assez proche de l'altermondialisme, Chirac (i) dénonce les « dangers » de la mondialisation. Seulement, (ii) il juge le mouvement « inéluctable » et, pourquoi pas, porteur de certains « avantages ». Aussi (iii) milite-t-il pour une mondialisation « maitrisée » (voir Mayaffre 2004 : 133-140)]

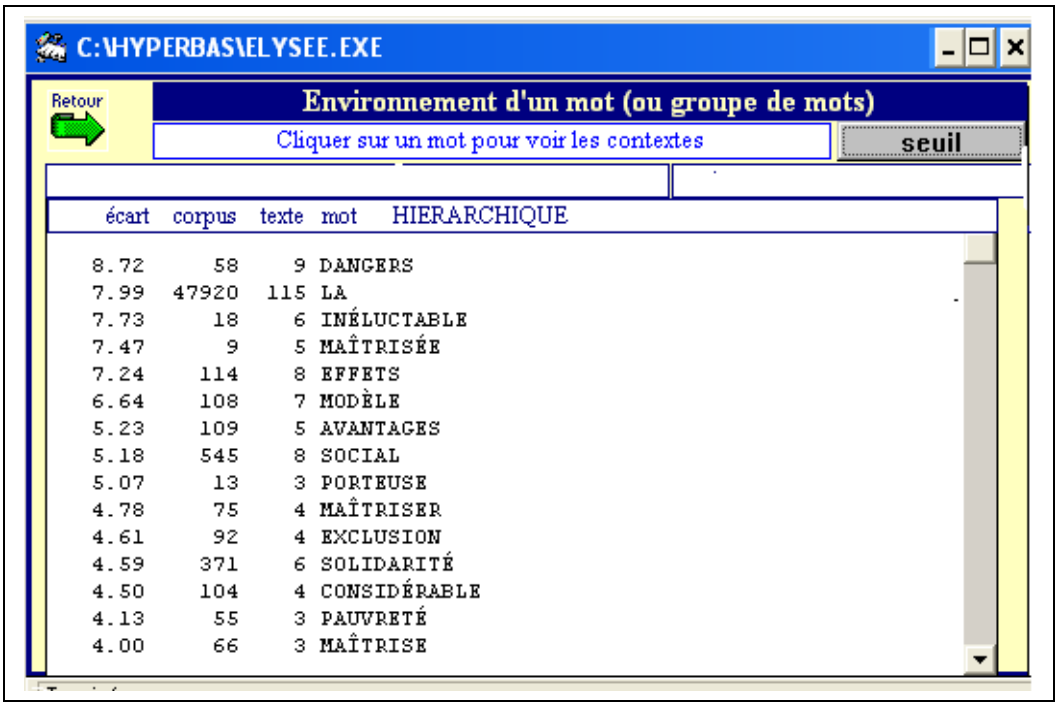


Illustration 3 : Environnement lexical du mot « mondialisation » dans le discours de J. Chirac

Bref, dans une concession décisive, [Adam 2006], un des meilleurs représentants de la linguistique textuelle traditionnelle, peut ainsi déclarer récemment devant les chercheurs en ADT :

> *…**la textualité doit résolument être pensée comme la combinaison de parcours linéaires et réticulaires.** [Adam 2006 : 5, souligné par l'auteur]*

Comme l'on sait que l'organisation du parcours linéaire a été le fait de la linguistique textuelle et des lectures oculaires depuis plusieurs lustres, l'on comprend que l'organisation du parcours réticulaire, désormais partie intégrante de la compréhension d'un texte, est laissée à la charge de l'approche assistée par ordinateur seule à même de formaliser des réseaux trans-phrastiques et a-séquentiels, à partir du moment où le texte est long et qu'il s'inscrit dans de gros corpus dont on prétend rendre compte[1].

1.2. Le corpus est un construit… qui construit

Les corpus ne sont pas des objets donnés mais des objets construits. Cette affirmation, qui n'est plus, espérons-le, à démontrer[2], n'est pas, elle non plus, le fait du tournant numérique. Elle prend cependant un tour particulier avec lui.

Si l'ordinateur dématérialise le texte en l'arrachant de son support physique habituel, il matérialise, délimite, organise –en un mot : construit– les corpus plus strictement qu'ils ne l'étaient auparavant. Dans les SHS, les corpus avaient parfois cessé, en effet, d'être des réalités pour devenir des potentialités. Selon l'exemple personnel cité, notre corpus papier était composé des discours de Maurice Thorez, que l'on savait exister dans les bibliothèques-archives les mieux documentées et que l'on pouvait lire à l'occasion ici ou là. Mais jamais il n'a pris la forme d'un objet autre qu'intellectuel. D'autre part, et conséquemment, son organisation était pratiquement nulle. Au mieux pouvait-on se prévaloir d'une hiérarchie chronologique dans la pile partielle de photocopies que l'on envisageait de faire et de quelques fiches de renvoi d'un texte à l'autre susceptibles de suppléer l'organisation informelle –l'anarchie ? – de notre mémoire.

Le numérique est jusqu'à nouvel ordre plus contraignant en matière de constitution et définitivement plus performant en matière d'organisation.

– *Constitution*. L'on ne pourra considérer, de droit comme de fait, comme appartenant aux corpus que les textes que l'on aura fait l'effort de *saisir* (dans son acception pleine mais d'abord physique) et que l'on pourra soumettre, effectivement, aux logiciels d'exploitation. Si l'esprit humain peut se satisfaire de potentialités, avantageusement ajustables au fil de la recherche, le système binaire des logiciels (oui/non) ne supporte que les choix définitifs et les traitements ne pourront s'opérer que sur des objets réellement constitués. Par là, *la clôture* des corpus est toujours contraignante dans le travail numérique, lorsque les chercheurs avaient tendance à élargir ou rétrécir au cours de leur étude, au gré de leur humeur, leur corpus d'étude. Dans les termes de [Pincemin et Rastier 1999], une certaine confusion, au moins une porosité, était souvent maintenue entre *corpus de travail* et *corpus de référence*. Aujourd'hui, de manière implacable, un texte fera partie ou non du corpus de travail. Le simple décompte par l'ordinateur des unités linguistiques du corpus, par exemple, ne peut supporter aucune ambiguïté quant à l'appartenance ou non d'un texte au corpus de l'analyse ; plus généralement, le traitement statistique de la lexicométrie opère nécessairement, selon les lois de la norme endogène, sur des corpus clos et réels.

Cette clôture du corpus –essentielle pour la rigueur de la démarche scientifique– va de pair avec la prétention de l'exhaustivité du traitement. Clos, délimité, le corpus numérique sera soumis *dans sa totalité* au même traitement systématique et exhaustif. Là encore, il en va de l'entêtement algorithmique des machines. La recherche d'un mot par exemple, puis de ses co-occurrents, dès lors qu'elle pourra se faire ici, pourra s'effectuer partout dans le corpus, rompant ainsi avec le caractère aléatoire, partiel et partial de l'attention humaine.

Enfin, cette exhaustivité du traitement prendra sa valeur seulement lorsqu'on aura indiqué que la taille des corpus numériques semble ne pas avoir de limite là où la mémoire humaine ne peut embrasser que des ensembles de quelques dizaines de textes. Sans assistanat numérique (moteur de recherche, indexation lexicale, navigation hypertextuelle, traitement quantitatif, tri alphabétique ou hiérarchique, concordanciers), il paraît difficile de prétendre rendre compte d'un

---

[1] Explicitement : « Nous avons, de toute évidence, besoin les uns des autres : tandis […] que nous mettons l'accent sur la définition des unités élémentaires, sur le traitement de la linéarité des textes, sur les enchaînements transphrastiques et sur la combinatoire d'unités de rangs de complexité supérieurs à la phrase, vos travaux insistent sur la structure non-séquentielle et réticulaire des textes. » [Adam 2006 : 4 ; propos tenu à la communauté ADT le 19 avril 2006, à Besançon, à l'occasion 8ème JADT].

[2] Voir, par exemple, la philosophie de la revue *Corpus*, notamment, dès le premier numéro [Mellet 2002], puis [Scheer 2004], [Mayaffre 2005-b] etc.

corpus de 100 discours politiques ; avec assistanat, il devient aisé de fouiller des corpus qui en comptent plusieurs milliers. Ce changement d'échelle de la taille des corpus, qui rend difficilement contournable les descriptions quantifiées, est en lui-même déterminant, d'autant que les traitements d'ADT se fixent comme objectif de combiner analyse globale [décompte systématique des unités, typologies des textes, classifications automatiques ; ceci sur de grands corpus] et analyse locale [retour au cœur des textes, pointages hypertextuels et repérages spatiaux des unités dans leurs contextes (le mot, la syllabe, la lettre dans la partie, le paragraphe, la phrase)]. Conçus pour cela, les logiciels défrichent et déchiffrent ; imposent au corpus un traitement synthétique et un traitement analytique, articulent, pour reprendre la terminologie de l'herméneutique, l'analyse du *tout* (en général grâce aux fonctions d'exploitations statistiques) et l'analyse des *passages* (en général grâce aux fonctions d'explorations documentaires) [1].

– *Organisation*. Si le numérique apporte une rigueur appréciable dans la constitution (entendons donc pleinement : *la saisie*) de gros corpus, il offre surtout une possibilité sans précédent de les organiser afin de mieux les interpréter. C'est ici que se trouve l'enjeu épistémologique le plus important de la philologie et/ou herméneutique numérique.

Le sens naît en/du contexte. La linguistique textuelle pose que celui-ci est minimalement le texte. Sans ignorer la rupture que cela constitue avec la tradition saussurienne orthodoxe, il apparaît aujourd'hui que cet élargissement de l'objet de la linguistique de la phrase au texte, pour être subversif, n'est pas suffisant. Car dans la recherche ou la construction du sens, aucun texte ne se suffit à lui-même.

Il s'agit-là de thèses inutiles à plaider sauf à remettre en cause les notions établies de co-texte, d'intertextualité ou de dialogisme et à ignorer quelques grands auteurs tel Bakthine.

Précisément, la linguistique de corpus telle que nous la concevons se propose de formaliser, autant que possible, cet au-delà du texte. Elle considère les corpus bien conçus comme des lieux nécessaires qui permettent d'objectiver le co-texte des textes qui les composent, c'est-à-dire, comme des réseaux sémantiques auto-suffisants (ce que ne sont pas les textes seuls). Mieux : elle considère avec [Rastier 1998 : 17] que « le corpus est la seule forme possible d'objectivation de l'intertexte » immédiatement nécessaire à l'interprétation des textes constituants. En un mot, les corpus numériques –par leur taille et leur organisation– doivent être élaborés et perçus comme des architextes sémantiques qui comprennent, en leur sein, les ressources textuelles nécessaires à leur compréhension/interprétation[2].

Nous avons effectivement pointé ailleurs ([Mayaffre 2002-b]) l'injustifiable inégalité de traitement entre les textes analysés (le corpus) et les textes mobilisés comme ressources interprétatives (l'intertexte ou, pour restreindre le propos, le co-texte). Quoique de même nature textuelle, les premiers font l'objet d'une approche scientifique (sélection, regroupement, traitement linguistique), les seconds interviennent, à discrétion dans l'analyse, sans autre précaution. C'est pour palier cette anomalie épistémologique que le numérique et les possibilités qu'il donne, doivent permettre de fondre autant que possible source et ressources textuelles au sein même du corpus.

Pour ne pas manquer la vocation que nous lui assignons, à savoir celle de matrice du sens, le corpus doit donc tendre vers la mise en forme de parcours sémantiques ou interprétatifs valides et fertiles ; parcours endogènes au corpus donc, dans lesquels, répétons-le, texte et co-texte ne sont pas discriminés et où les ressources interprétatives se trouvent internalisées.

Pour cela, nous avons insisté sur la dimension *réflexive* que les corpus gagnent à avoir. En miroir, les textes du corpus doivent s'éclairer mutuellement ; se *réfléchir* les uns les autres ; chacun d'entre eux constituant le co-texte immédiat de tous, et l'ensemble, l'intertexte de chacun. Ainsi par exemple, l'étude du discours de Jacques Chirac qui a été entreprise [Mayaffre 2004] est passée par un corpus qui comprenait outre les textes du président actuel, ceux de ses prédécesseurs à l'Elysée. Les discours de de Gaulle, Pompidou, Giscard et Mitterrand constituaient à nos yeux l'intertexte générique et l'intertexte historique du discours chiraquien. Le corpus comprenait aussi

---

[1] On ne saurait trop insister ici sur la puissance et la souplesse des ordinateurs pour décomposer/recomposer un tout en parties. Dans une mise en abîme impressionnante, l'utilisateur paramétrera dans sa recherche (i) l'unité à rechercher (le segment, le syntagme, le mots ou co-occurrents, les lemmes ou les codes grammaticaux, la chaîne de caractères, la syllabe ou la lettre) et (ii) la largeur de la fenêtre textuelle qui lui semblera pertinente pour la contextualisation (le corpus dans sa totalité, les textes, les parties, le paragraphe, la phrase, le début de ligne, la fin du vers, etc.).

[2] Se risquera-t-on ainsi à préciser que le corpus devient alors l'objet nécessaire et maximal –comment imaginer un objet constitué ou empirique plus important ? – d'une linguistique aboutie ?

les discours de Lionel Jospin car ils semblaient constituer, pendant la période de la cohabitation, le co-texte politique, immédiat et incontournable, des propos du président. Dès lors, les mots de l'insécurité de Chirac, par exemple, n'ont pris sens qu'en considérant ceux que prononçait, en contrepoint, le Premier ministre durant la même période.
D'un point de vue technique, précisons simplement que la *réflexivité du corpus*, c'est-à-dire, au fond, la mise en dialogue des textes constituants, est assurée avant tout par les vertus de l'hypertextualité. Celle-ci semble être une solution puissante pour formaliser la notion d'architextualité de [Genette 1979] et rendre possible cette *sémantique de l'intertexte* que réclamaient les auteurs du [*Cahier de praxématique* 1999]. Les logiciels d'analyse textuelle sur le marché considèrent en effet les corpus comme de vastes hypertextes : toutes les unités sont indexées et liées les unes aux autres. Le mot « délinquance », par exemple, trouvé dans la bouche de Chirac le 14 juillet 1996, renvoie non seulement à toutes les occurrences du mot dans le discours du président, mais dans ceux de Jospin. Et cette mise en résonance des textes, impossible à imaginer manuellement, est systématisée, grâce à l'hypertextualité et au traitement statistique contrastif, pour *toutes* les unités dans *l'ensemble* du corpus.
Quoique utopique l'idée de corpus réflexifs, c'est-à-dire la prétention d'internaliser les ressources textuelles interprétatives au sein de gros corpus dûment constitués, a été reprise par [Guilhaumou 2002 : 40], [Rastier 2005 : 31-32] et [Adam 2006 : 16], chacun insistant sur la dimension philologique et/ou herméneutique de la proposition. Car cette propriété des corpus, comme d'autres propriétés sur lesquelles il est impossible de revenir, revient à admettre qu'un « **moment philologique** » [Adam 2005 : 83 souligné par l'auteur] doit présider à leur constitution : l'acte interprétatif devant être pressenti au moment de la sélection et de l'organisation des textes en corpus. Au-delà de l'inévitable circularité de la démarche (cf. *infra 2.2,* la question du cercle herméneutique), il s'agit de circonscrire le problème épineux du « point de vue » au seul geste inaugural de la recherche (le corpus comme un « point de vue ») pour mieux objectiver ensuite, dans le reste de l'analyse, l'interprétation.

2. Vers un contrôle de l'interprétation

L'enjeu des sciences du texte est moins d'administrer la preuve que de contrôler l'interprétation. Hors de l'obscurantisme théologique, il faut admettre en effet que les textes, et les corpus qui en informent le sens, n'ont point de Vérité mais de multiples compréhensions. Selon une pensée attribuée à Foucault, la vérité d'un texte est d'abord et seulement ce qu'on dit de lui, et déjà Chladenius remarquait que, loin de la stricte intentionnalité des auteurs, « l'on peut, lorsqu'on cherche à comprendre leurs écrits, former des pensées qui n'étaient pas venues à l'esprit de l'auteur » [Chladenius cité par Szondi 1989 : 32].
Seulement, sauf à verser dans un subjectivisme débridé et une interprétation divinatoire, « ce qu'on dit des textes » et ces « pensées » qu'il est permis d'avoir à propos d'eux, doivent être étayés, vérifiables, contrôlés. Cela passe, comme indiqué, par une composition/organisation *ad hoc* des corpus, cela passe aussi par la prise en considération rigoureuse du matériel linguistique qui les constitue.

2.1. Une herméneutique matérielle

L'herméneutique numérique est une herméneutique matérielle ; pas seulement par conviction mais par nécessité. Ou plutôt : avec le numérique l'évidence devient nécessité.
Pour les raisons techniques indiquées plus haut, la machine en effet ne saurait embrasser le texte autrement que par sa matière. Sauf à renverser le procès de la démarche et s'illusionner sur les possibilités de l'intelligence artificielle, l'ordinateur ne peut donner accès au sens d'un texte sans appréhender sa lettre ; il ne saurait aborder son esprit sans traiter (« saisir », « implémenter », « digitaliser », « numériser ») sa matière.
Concrètement, du côté de la linguistique quantitative, se retrouve ici l'objection la plus pertinente de [Tournier 1985 et 1987] contre les lemmatisations aveugles et toute forme d'analyse lexicométrique reposant sur un traitement linguistique liminaire du corpus. Lemmatiser un texte, c'est ramener son vocabulaire (particulier, historique, idéologique) à un lexique (universel, intemporel) : c'est plaquer sur des textes historiques un sens préalable (celui canonique du

dictionnaire), là où l'analyse prétendait justement déconstruire/reconstruire froidement les textes pour faire percer le sens sous la surface matérielle, graphique –supposée neutre– du corpus.[1]
Se retrouve, aussi, ici, la critique la plus forte de la linguistique textuelle adressée à l'analyse du discours, qui, si elle n'y prend garde, « manque le texte en tant que tel » [Sarfati 2003 : 432] ; critique que l'on peut généraliser.
Manquer le texte pour l'analyse du discours, c'est prendre en considération les conditions de production des textes et négliger les productions elles-mêmes. Manquer le texte, pour l'herméneutique traditionnelle, c'est prétendre toucher l'âme des textes en négligeant leur chair. Manquer le texte pour la rhétorique, par exemple, c'est s'éblouir sur quelques fleurs de langage ou figures de style remarquables, lorsque la matérialité du texte, dans son ensemble, participe de l'éloquence du discours.
Chevillée donc à la matière textuelle, sûre de la description formelle des corpus, l'herméneutique numérique ne prétend certes pas produire des interprétations infalsifiables mais entend toujours s'appuyer sur des unités linguistiques attestées de textes établis. C'est en ce sens qu'elle peut se revendiquer de Peter Szondi et de son herméneutique critique ; c'est en ce sens que l'on parle d'une herméneutique philologique. Les parcours interprétatifs sont toujours sujets à caution[2], mais la trajectoire de ceux de la philologie et/ou herméneutique numérique a l'avantage d'être solidement inscrite dans la bonne direction grâce à son décisif et premier mouvement : par la prise en compte nécessaire, systématique et exhaustive, des matériaux linguistiques (lettres et syllabes, formes graphiques et lemmes, code grammaticaux et enchaînements syntaxiques, segments répétés, expressions, cooccurrents, collocations micro-distributionnelles, réseaux lexicaux, concordances phrastiques, contextes paragraphiques, etc.) des textes.

2.2. Cercle herméneutique et démarche inductive

La conséquence la plus directe de l'approche matérielle de la philologie et/ou herméneutique numérique est, nous semble-t-il, le caractère à dominante inductive de la démarche.
Les procédures de l'herméneutique, et peut-être celles de l'acquisition de la connaissance, sont prisonnières d'un *cercle* [Cf. la plupart des auteurs qui ont traité le sujet et particulièrement Schleiermacher]. La première façon d'évoquer ce cercle a été rappelée dès la note liminaire : philologie et herméneutique apparaissent attachées dans une relation sans commencement ni fin : l'établissement d'un texte passe par sa compréhension profonde c'est-à-dire son interprétation, et l'interprétation ne peut reposer que sur un texte solidement établi. La seconde façon d'évoquer ce cercle est plus classique dans la littérature sur l'herméneutique. Elle souligne comment l'analyse du tout et celle des passages se trouvent liées dans un rapport sans issue. Le passage ne pouvant être compris que dans/par l'ensemble, et l'ensemble ne pouvant être construit qu'à partir de la compréhension des parties.
Précisément, Peter Szondi citant plusieurs fois Heidegger, nous invite à renoncer à l'idée d'échapper à cette circularité de la compréhension, de l'interprétation et de la connaissance, et pose que « l'essentiel […] n'est pas de sortir du cercle, mais d'y entrer de la bonne manière » [Heidegger cité sous des formes différentes par Szondi 1989 : 10 et 105].
Pour notre part, l'on entre dans le cercle herméneutique, ou en tout cas dans le corpus, *par le bas*. Dans les termes que [Williams 2005] reprend à [Tognini-Bonnelli 2001], nos études sont *corpus-driven* (versus *corpus-based*) et la démarche *buttom–up* (versus *top-down*). La linguistique de corpus pose que le corpus n'est pas l'outil de la recherche (un corpus-ressources documentaires, un corpus-base de données, un corpus-échantillon représentatif de la langue) mais son objet vivant et dynamique ; il est non pas le réceptacle d'un sens déjà là mais sa matrice ; non pas une chose que l'on interroge, mais une chose qui nous interroge.
Dès lors, si le corpus construit un sens que l'on cherche à appréhender, si c'est lui qui, une fois constitué, conduit objectivement l'analyse, la meilleure démarche est celle qui permet de faire

[1] On objectera néanmoins à Maurice Tournier sa propre critique : la forme graphique qu'il réifie et semble considérer comme neutre est elle-même arbitraire, historique, conventionnelle. Le texte graphique n'est, guère plus qu'un texte lemmatisé, un texte objectif ou naturel. (Cf. [Mayaffre 2005-a]). Grâce à la performance des lemmatiseurs/étiqueteurs, et malgré leurs erreurs résiduelles, la surface lemmatisée ou grammaticalisée du texte n'est, aujourd'hui, guère moins contestable ou arbitraire que celle d'un texte brut.
[2] « Il est certain que l'on ne peut pas simplement biffer la part de subjectivité, et même d'affinité dans la démarche de la compréhension » [Szondi 1989 : 117-118].

remonter l'information du tréfonds, afin de nourrir le plus objectivement possible nos interprétations.
Cette démarche à dominante inductive est rendue cohérente par les contraintes du numérique telles qu'évoquées précédemment. L'ordinateur décompose ses objets en plus petites unités sémiotiques. Et un corpus est pour lui d'abord constitué de lettres concaténées, de blancs et de ponctuation, d'octets et de bits. Si ces unités sont ensuite combinées, reliées, contextualisées (voire interprétées comme dans le cadre de la lemmatisation), la description comme l'interrogation numériques du texte s'appuieront sur ces signaux informatiques premiers et minimaux.
Cette posture inductive est non seulement essentielle dans l'acte descriptif qui précède l'interprétation, mais dans la démarche interprétative elle-même. De fait, en partant d'en bas, les logiciels peuvent décrire des gros corpus avec la minutie, la systématicité et l'exhaustivité mentionnées. Ils outillent donc le chercheur dans sa recherche d'indices objectifs. Mais, par les lectures complémentaires qu'ils proposent et les visions alternatives qu'ils donnent des textes (cf. *supra*), les logiciels d'ADT doivent surtout interroger différemment l'herméneute, loin d'hypothèses de travail, imposées par en haut et trop contraignantes. Car la plus-value attendue de l'ADT et de la philologie et/ou l'herméneutique numérique est avant tout une plus-value heuristique. Il s'agit de retourner la démarche hypothético-déductive dont l'usage apparaît trop dangereux dans les sciences humaines [cf. Mayaffre 2002-a], en faisant émerger, du corpus même, des hypothèses objectives de travail sur lesquelles on se met à réfléchir. Il s'agit de refuser le risque de projeter dans les textes un questionnement surplombant et un sens préalable, pour se laisser interroger par eux sans *a priori* et sans tabou.
Mieux que contrôler les parcours interprétatifs en se donnant les moyens de décrire puis de vérifier nos (hypo)thèses sémantiques, l'ADT se propose d'objectiver l'élaboration desdites hypothèses. Contrôler les conditions d'émergence des hypothèses apparaît ainsi comme le suprême objectif de l'herméneutique numérique, pour une meilleure maïeutique du sens.

Conclusion

Revenons pour conclure à l'élémentaire. Le moyen le plus simple de souligner l'importance de la révolution numérique dans la perception des textes, des corpus et dans les pratiques interprétatives est sans doute de montrer l'amplification décisive qu'elle représente avec le meilleur de la philologie et/ou herméneutique traditionnelle.
N'a-t-on pas souligné dans les humanités, l'apport scientifique de l'invention de la glose et des notes infrapaginales. Il s'agissait, tout à coup, d'enrichir le texte d'un paratexte et de renvoyer le lecteur à des références bibliographiques ou à des sources, bref à d'autres textes. La révolution numérique permet de rendre effective ce système d'enrichissement et de références et de transporter, sur le champ, le lecteur au-delà de son document d'origine. Dans l'édition électronique des *Fleurs du mal* qu'a entreprise [Viprey 2002], par exemple, les différentes éditions de l'œuvre de Baudelaire sont instantanément consultables ainsi que des dictionnaires, des graphes ou des index. Bien sûr, en cela, les pratiques numériques redéfinissent la conception du texte, en violent les frontières physiques (classiquement le livre), et renoncent définitivement à le percevoir comme une monade.
N'a-t-on jamais apprécié de pouvoir entrer dans une œuvre par l'index des noms de personnes, de lieux ou de notions ? Les corpus numériques et les logiciels d'ADT, dans leur plus simple appareil, se proposent, comme première étape, d'indexer l'ensemble des mots qui seront autant d'entrées au cours de la recherche. Outre l'amplification du travail d'indexation, cette généralisation signifie d'un point de vue épistémologique que les *a priori* quant aux mots jugés pertinents disparaissent. Par là, c'est donc une interrogation non bornée que l'on s'autorise, jusqu'à un renversement du système hypothético-déductif qui implique, toujours, une lecture orientée pour des interprétations convenues.
Enfin, n'a-t-on jamais navigué, à la recherche du sens, dans un dictionnaire ou une encyclopédie entre plusieurs articles *via* le système de renvois thématiques ? La conception numérique des corpus multiplie ces passerelles ; elle n'est que passerelles. Elle rend industrielle et, pour tout dire, enfin opératoire, l'artisanat dérisoire des renvois manuels. Par simple clic, la navigation hypertextuelle fait passer le lecteur, au sein du corpus, d'un mot à l'autre (de tous les mots à tous les autres), d'un texte à l'autre, d'un thème à l'autre. Les corpus textuels par leur taille et leur structure peuvent être perçus comme de gros architextes : leur lecture hypertextuelle et leur traitement statistique sont la condition de leur exploitation en tant que tels. L'enjeu apparaît alors

aussi simple qu'évident : faciliter, organiser, contrôler, mieux qu'auparavant, la contextualisation des mots, des phrases et des textes constituants, sans laquelle aucune interprétation scientifique n'est envisageable.

BIBLIOGRAPHIE

ADAM, J.-M. 1999. *Linguistique textuelle. Des genres de discours aux textes*, Paris, Nathan.

ADAM, J.-M. 2005. Les sciences de l'établissement des textes et la question de la variation, in J.-M. Adam et U. Heidmann (éds.), *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité*, Genève, Slatkine, pp. 69-92.

ADAM, J.-M. 2006. Autour du concept de *texte*. Pour un dialogue des disciplines de l'analyse de données textuelles, in *JADT 2006* [texte en ligne sur *Lexicométrica (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf)*].

Cahiers de praxématique 1999. « Sémantique de l'intertexte », n°33.

CHARAUDEAU, P. et MAINGUENEAU, D. (sous la dir.) 2002. *Dictionnaire d'analyse du discours*, Paris, Seuil.

DÉTRIE, C., SIBLOT, P., VÉRINE, B. 2001. *Termes et concepts pour l'analyse du discours. Une approche praxématique*, Paris, Champion.

GENETTE, G. 1979. *Introduction à l'architexte*, Paris, Seuil.

GUILHAUMOU, J. 2002. Le corpus en analyse de discours. Perspective historique, *Corpus,* 1, pp. 21-49.

HEIDMANN, U . 2005. Comparatisme et analyse de discours. La comparaison différentielle comme méthode, in J.-M. Adam et U. Heidmann (éds.), *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité*, Genève, Slatkine, pp. 99-116.

MAYAFFRE, D. 2002-a. L'Herméneutique numérique, *L'Astrolabe. Recherche littéraire et Informatique* (http://www.uottawa.ca/academic/arts/astrolabe/).

MAYAFFRE, D. 2002-b. Les corpus réflexifs : entre architextualité et intertextualité, *Corpus*, 1, pp. 51-70 (http://revel.unice.fr/corpus/document.html?id=11).

MAYAFFRE, D. 2004. *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la V*ème *République*, Paris, Champion.

MAYAFFRE, D. 2005-a. De la lexicométrie à la logométrie, *L'Astrolabe. Recherche littéraire et Informatique* (http://www.uottawa.ca/academic/arts/astrolabe/).

MAYAFFRE, D. 2005-b. Les corpus politiques : objet, méthode et contenu. Introduction, *Corpus,* 4, pp. 5-19.

MELLET, S. 2001. Corpus et recherches linguistiques : introduction, *Corpus*, 1, pp. 5-13.

PINCEMIN, B. et RASTIER, F. 1999. Des genres à l'intertexte, *Cahiers de Praxématique*, 33, pp. 83-111.

RASTIER, F. 1998. Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage, *Langages*, 129, pp. 97-111.

RASTIER, F. 2001. *Arts et sciences du texte*, Paris, Puf.

RASTIER, F. 2005. Enjeux épistémologiques de la linguistique de corpus, in G. Williams (éd.), *La linguistique de corpus*, Rennes, Pur, pp. 31-45. [En ligne sur *Texto !* (http://www.revue-texto.net/Inedits /Rastier/ Rastier_Enjeux.html)]

SCHEER T. 2004. Le corpus heuristique : un outil qui montre mais ne démontre pas, *Corpus*, 3, pp. 153-193.

SZONDI, P. (trad. 1989). *Introduction à l'Herméneutique Littéraire. De Chladenius à Schleirmacher*, Paris, Cerf.

TOGNINI-BONNELLI, E. 2001. *Corpus Linguistics at Work*, Amsterdam, John Benjamin's Publishing.

TOURNIER, M. 1985. Sur quoi pouvons-nous compter ? Réponse à Charles Muller, in *Études de philologie et de linguistique offertes à Hélène NAIS*, *Verbum* (numéro spécial), Presses universitaires de Nancy.

TOURNIER, M. 1987. *La réduction : principe de lexicométrie politique*, brochure de URL "Lexicométrie et textes politiques", 14 pages.

VIPREY, J.-M. 2002. *Analyses textuelles et hypertextuelles des* Fleurs du mal, Paris, Champion.

VIPREY, J.-M. 2005. Philologie numérique et herméneutique intégrative, in J.-M. Adam et U. Heidmann (éds.), *Sciences du texte et analyse de discours*, Genève, Slatkine, pp. 51-68.

WILLIAMS, G. 2005. Introduction, in G. Williams (éd.), *La linguistique de corpus,* Rennes, Pur, pp. 13-18.

OUI OU *NON* À LA CONSTITUTION EUROPÉENNE : L'ÉLOQUENCE DU FORUM [1]

Jessica MANGE, Pascal MARCHAND, André SALEM
IUT Caen, LERASS / IUT Toulouse 3, SYLED – CLA2T / Université Paris 3

SOMMAIRE

Résumé : *Les analyses médiatiques dominantes sur le référendum pour le traité établissant une constitution pour l'Europe peuvent être résumées de la façon suivante (Voir Le débat, n°136 ; Les cahiers du CEVIPOF, n°42) :*

- *Il s'est agi d'un débat citoyen exceptionnel et passionné ;*
- *Le résultat a été marqué par des préoccupations plus « nationales » qu'européennes, pour deux raisons :*

**une campagne monothématique menée par les tenants du « non » et concentrée sur des questions sociales ;*
**un « cadrage » médiatique de la campagne plus national qu'européen.*
Ces analyses dominantes présentent un certain nombre de paradoxes (les médias sont dans la position d'avoir favorisé la parole du « oui », tout en ayant développé une actualité sociale aboutissant au vote « non ») et de zones d'ombre (la plupart des analystes ne se donnent pour objet que de comprendre le vote « non »).
Notre position est de nous démarquer de ces analyses dominantes qui, soit en restent aux discours des élites censés refléter l'état « culturel » du débat à un moment donné, et dont on suppose l'application homothétique chez les citoyens, soit infèrent les opinions de ces mêmes citoyens de sondages d'opinion dont les questions - et leur formulation - sont largement guidées par les hypothèses posées a priori des enquêteurs ou des commanditaires.
Nous cherchons donc à vérifier, dans la conversation elle-même, la pertinence des analyses qui précèdent. Plusieurs méthodes relevant de l'analyse des données textuelles permettent de retracer la chronologie du débat, puis d'en dégager les arguments contradictoires et leur évolution.
Si l'on reproche parfois à l'approche informatique des textes un certain réductionnisme, force est de constater qu'elle permet ici, au contraire, d'aller au-delà des positionnements convenus et d'entrevoir des complexités qui dépassent les apparences commodes.

1. Caractéristiques du corpus

On analyse un corpus rassemblant les contributions de participants à un forum de discussion ouvert sur un site web français (telerama.fr) durant les six semaines qui ont précédé le scrutin du 29 mai 2005. Deux mille deux messages ont été « publiés » sur le forum entre le 22/04/05 et le 22/06/05 (415 pseudonymes différents). Soixante-quinze messages ont également été envoyés pour être publiés dans les numéros précédant le 29 mai 2005.

Les messages recueillis sur le forum sont généralement assez courts, avec une moyenne de 74 mots. Rares sont ceux qui dépassent les 500 mots, mais les plus gros peuvent exceptionnellement atteindre 2000 mots. Le corpus total représente donc 326223 occurrences correspondant à 19092 formes lexicales différentes. Le style dominant, tout en étant une écriture de l'oral, respecte les règles principales de l'écrit canonique : les messages sont majoritairement écrits en minuscules accentuées, avec peu de problèmes de caractères non-reconnus.

L'analyse de ces discours a eu recours à des logiciels relevant de la statistique textuelle[2]. Si le but ultime de l'analyse du corpus est bien d'en produire une lecture interprétative, il s'agit ici de proposer une conduite de détour (voir Salem) dont chaque étape implique des opérations formelles qui, pour être automatisées, n'en impliquent pas moins différents niveaux interprétatifs.

[1] Cette communication présente une synthèse de travaux publiés dans les actes des *Huitièmes Journées Internationales d'Analyse des Données Textuelles* et à paraître dans la revue *Mots*.
[2] Voir, par exemple, Lebart & Salem (1994), Marchand (1998, 2002).

Nous avons proposé de définir deux lignes interprétatives, que nous avons nommées horizontale et verticale, et que résume le tableau ci-dessous (voir Marchand à paraître).

Interprétations	**Codage**	**Analyse**	**Interprétation**
Horizontales	Segmentation, réduction	Concordances, cooccurrences…	Sémantique des formes
Verticales	Partition	Analyse des distances, des correspondances, classifications…	Sémantique des profils (plans factoriels, arbres, classes)

Les interprétations au niveau du codage peuvent être ramenées aux décisions suivantes :

- Quelques abréviations courantes (pb ou pbm, qq ou qlq, bcp, qd, M., Mme) ont été remises à leur forme canonique.
- Les fautes de frappe ou d'orthographe ont été corrigées, et notamment les fautes d'accents (a/à, la/là, traite/traité, prive/privé).
- Les (rares) dysorthographies visiblement volontaires (*pôvre, ceusses, référendome, zommes politiques, môssieur*) ont été conservées.
- Les fautes de grammaire n'ont pas été corrigées. En revanche, pour une partie des résultats suivants, nous avons eu recours à une lemmatisation partielle ramenant à la forme infinitive la plupart des formes verbales. La majorité des fautes d'accord – genre, nombre, participes passés ou confusions entre les formes du futur et du conditionnel, par exemple - ont ainsi été neutralisées.

Sur cette base, trois questions se sont posées à nous, impliquant des partitions et des analyses spécifiques, mais se situant toutes dans une lecture « verticale » du corpus.

2. Comment évolue le vocabulaire des participants tout au long du débat ?

Nous avons partitionné le corpus en six semaines (de la semaine 17 à la semaine 22 du calendrier 2005) :

- du 22 avril au 1er mai (161 messages)
- du 2 mai au 8 mai (268 messages)
- du 9 mai au 15 mai (512 messages)
- du 16 mai au 22 mai (464 messages)
- du 22 mai au 29 mai (560 messages)
- du 30 mai au 3 juin (36 messages auxquels s'est ajouté un message du 22 juin) : cette semaine suit la consultation électorale.

Il devient possible alors de construire un tableau croisant, en lignes les 2703 formes lexicales apparaissant plus de dix fois dans le forum, et en colonnes les six semaines analysées. L'analyse factorielle des correspondances (AFC) permet de rendre compte des distances entre les lignes, d'une part, et les colonnes, d'autre part.

Le premier axe de l'AFC restitue intégralement la chronologie du corpus. Cela traduit le fait, souvent constaté dans le cas de l'étude des séries textuelles chronologiques ou simplement de corpus longitudinaux, que le vocabulaire employé par les participants évolue progressivement dans le temps[1].

Après identification du schéma général de l'évolution progressive du vocabulaire au fil des semaines, il devient possible de caractériser chacune des parties (ou groupe de parties) par le vocabulaire et les séquences qu'elle privilégie et par celui qu'elle évite. Le lexique spécifique de chaque semaine est ainsi mis en relation avec l'actualité politique et médiatique du débat.

Pour résumer les grandes lignes de l'évolution chronologique, il est possible de rechercher les formes lexicales et segments répétés non seulement les plus significatifs de chacune des six périodes, mais également celles qui traduisent au mieux la chronologie de l'ensemble (Salem, 1993). Nous considérerons ainsi la distribution des formes les plus significatives de la première et des deux dernières semaines.

Les unités textuelles les plus employées dans les deux premières semaines se révèlent être les formes qui réfèrent en principe à l'objet du débat : *constitution, traité, traité constitutionnel, l'union,*

[1] Pour des compléments sur les *séries textuelles chronologiques*, on consultera [Salem 1993].

politique. La première période s'organise donc sur la question générale de l'objet « *constitution* », de sa description (*notion, articles, règlement*) et de ses conséquences (*concurrence, SIEG et service public*).
Ces formes sont nettement moins utilisées dans les semaines qui correspondent à la fin du débat sur le forum. De même, la distribution des segments répétés les plus longs (supérieurs à 10 formes), qui renvoient à de longues citations, révèle que les internautes y ont souvent eu recours dans les premières semaines du débat mais qu'ils diminuent fortement dans les dernières parties :

… pour éviter que le fonctionnement du marché intérieur ne soit affecté …
… les états membres s'efforcent de procéder à la libéralisation des …
… qui correspondent au remboursement de certaines servitudes inhérentes à la notion …
… toute personne a le droit de travailler et d'exercer une …
… un état membre peut être appelé à prendre en cas de …
… des citoyens de l'union, au nombre d'un million au …

Dans ces dernières périodes, on note au contraire une plus grande utilisation des termes : *tu, vous, dimanche, docteurs, patient, remède*, que l'on peut interpréter comme la marque d'un discours plus polémique, plus centré sur l'imminence de l'échéance électorale et dans lequel les enjeux principaux ont déjà été décrits. Le débat s'est déplacé vers l'affrontement de personnes (les « *tenants* », « *partisans* » ou « *défenseurs* » deviennent « *nonistes* » et « *ouistes* », et deviendront « *le camp* » après le vote), qui confinera même à l'insulte dans les quelques messages suivant immédiatement le scrutin : *mépris, mauvaise foi, tricheurs, imposteurs, manipulateurs*.
Ainsi peut-on confirmer la tonalité fortement passionnée du débat à propos de la Constitution européenne, tout en notant la montée en puissance progressive du vocabulaire exprimant cette passion.

3. Quels sont les arguments du débat ?

Nous avons cherché à définir statistiquement des classes d'arguments. La méthode implique ici une partition très différente et repose sur la classification hiérarchique descendante (CDH) d'un tableau lexical croisant le lexique (lemmatisé) avec des unités de contexte (Reinert, 1990) : partant du corpus intégral, on cherche à définir, de façon itérative, des classes lexicales statistiquement indépendantes. Il s'agit ensuite de voir si ces classes d'arguments peuvent être corrélées avec le « oui » ou le « non ».
Un codage des messages a donc été effectué manuellement : seuls 71 messages (3,5%) n'ont pas pu être clairement attribués ; les autres se répartissent de façon équilibrée en 1000 « pro-oui » (48,1%) et 1006 « pro-non » (48,4%). Rappelons que ce codage n'intervient pas dans l'analyse du corpus elle-même mais en constitue une illustration possible.
Six classes lexicales sont issues de la CDH : deux sont corrélées avec le « oui », deux avec le « non » et deux ne sont corrélées ni avec l'un, ni avec l'autre. Nous donnons ci-dessous des termes significatifs de chacune des classes ainsi que des exemples de réponses significatives (également extraites de façon automatique).
Deux classes lexicales, issues de la même classe-mère, ne sont pas corrélées avec les codes *oui* et *non*. Il s'agit, d'une part de commentaires sur le forum de Télérama (avec les termes : *argument, Télérama, forum, je, lire, message, lecteur, merci, lire, éditorial, médias, débat, journal, bravo*) et d'autre part d'arguments techniques portant sur la procédure du référendum (*traité, question, poser, Nice, vote, renégociation, texte, négociation, approuver, point, référendum, emporter, ratifier, politique, revenir*).
Exemple de message de la classe 4 :

> *« …un grand bravo à Télérama pour, d'une part, reconnaître dans ses colonnes* que 65 % du **courrier reçu viennent** de **partisans** du non, alors que **Télérama** … ».

Exemple de message de la classe 2 :

> « on ne nous a pas **demandé** notre avis sur les **traités actuels** ? **Erreur** */elatorn26/*, le **traité de Maastricht** a été **approuvé** par **référendum**. Quant aux **traités** ultérieurs, ils ont été **ratifiés** par un parlement **français** qui, il me semble, est élu et donc **responsable** devant ses électeurs ».

Ces deux classes, non liées aux opinions, ne présentent pas un intérêt majeur pour notre étude et nous n'en dirons pas davantage.

Deux classes lexicales sont corrélées avec le « non ».
Nous pourrions nommer la première « discours politique et citations du Traité ». Les termes significatifs de cette classe sont en effet : *service, public, concurrent, libre, fausser, entreprise, économie, article, général, intérêt, marché, travail, protection, emploi, privé* ainsi que la modalisation (*non, si, oui, pas, trop, ne, point, déjà, de-manière, pouvoir., donc, certes, cependant, falloir., vouloir., de-toute-facon, pour, surtout, mais, sur, parce-qu<, à-travers, en-cas, d'-abord, aujourd'-hui, il-me-semble*).
Exemple de message de la classe 3 :

> « … **article** I.3.3. l'union oeuvre pour le **développement durable fondé** sur une **croissance économique équilibrée** et une stabilité des **prix,** une **économie sociale** de **marché hautement compétitive.** Les **services publics économiques** restent **soumis** aux **règles** de la **concurrence,** III.166, et à la limitation des **aides publiques,** III.167 … »
> « … en-plus, l' **article** III.167 **interdit** toute **aide publique** aux **entreprises publiques:** sont incompatibles avec le **marché intérieur** les **aides accordées** par les états membres sous quelques **formes** que ce soit qui **fausse** ou qui **menacent** de **fausser** la **concurrence** en **favorisant** certaines **entreprises** ou certaines productions ».

La deuxième classe lexicale corrélée au « non » fait également largement appel au Traité lui-même, mais davantage en référence aux « cadres institutionnels de l'Europe » : *parlement, union, commission, membre, Etat, conseil, charte, compétent, droit, européen, initiative, OTAN* ainsi que des adjectifs et adverbes (*fondamental, législatif, exécutif, européen, national, coopératif, respectif, militaire, universel, étranger, antérieur, international, soumis, significatif, nécessaire, parlementaire, fédéral, progressif, défini, nouveau, précédent, direct*).
Exemple de message de la classe 6 :

> « …les **dispositions** de la **présente charte** qui **contiennent** des principes peuvent être **mises** en **oeuvre** par des **actes législatifs** et **exécutifs** pris par les **institutions, organes** et organismes de l'**union** et par des **actes** des **états** membres … »

Ces deux classes lexicales, issues de la même classe-mère, ce qui confirme leur cohérence statistique, renvoient donc à un débat argumenté sur la base du contenu même du traité et de son analyse. Les internautes proposent ici une lecture interprétative des conséquences du texte.
Enfin, deux classes lexicales, encore une fois issues de la même classe-mère, sont corrélées avec le « oui ».
Nous avons proposé de nommer la première « discours philosophiques (concepts et valeurs) ». Les termes significatifs de cette classe sont en effet : *monde, nous, histoire, vivre, mondial, rêve, Europe, pays, guerre, solidarité, pauvre, petit, vie, chômage, modèle, jeune(s)* ainsi que des marqueurs de lieux et de pays.
Exemple de message de la classe 1 :

> « …nous n'avons jamais vraiment voulu directement **aider** les **pays** plus **pauvres** que nous à se développer. Où en serait (*sic*) l'**Espagne** et le **Portugal** aujourd'hui sans l'**Europe** ? **L'Europe** est le **seul exemple récent** de **réelle solidarité** entre les **pays** les plus **riches** et les plus pauvres… ».

La deuxième classe lexicale corrélée avec le « oui » est celle des acteurs politiques : *gauche, le Pen, droite, Fabius, Villiers, Chirac, extrême, parti, PS, Jospin, présidentielle, Besancenot, Buffet, Hollande, tour*. On y trouve également des adjectifs et adverbes, mais d'une tout autre tonalité que dans la classe 3 : *souverainiste, populeux, électif, trotskiste, populiste, xénophobe, frileux, électoral, partisan, prochain, généreux, majoritairement, second, majoritaire, quasiment*.
Exemple de message de la classe 5 :

> « …mais quand dans le **camp** du non on **retrouve** le **Pen,** les **trotskistes,** de **Villiers,** Boutin, **Fabius** et Mégret on se dit que la **démagogie** est plus importante dans le **camp** du non… ».

Si les deux classes corrélées avec le « non » présentaient une certaine cohérence, les deux classes corrélées au « oui » semblent relever de registres différents. Quoi de commun, en effet, entre l'idéal européen et les acteurs de la politique française ?
Le fait que les arguments du *Oui* et du *Non* soient distingués dans une classification automatique calculée indépendamment de ce codage nous autorise à approfondir leur spécificité.

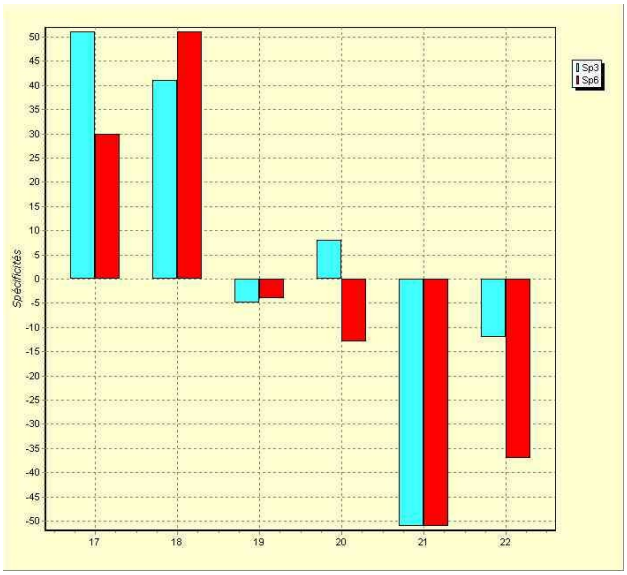
4. La spécificité et la chronologie des arguments du *oui* et du *non*

La recherche des formes et segments répétés spécifiques du *Oui* et du *Non* confirme la CDH :

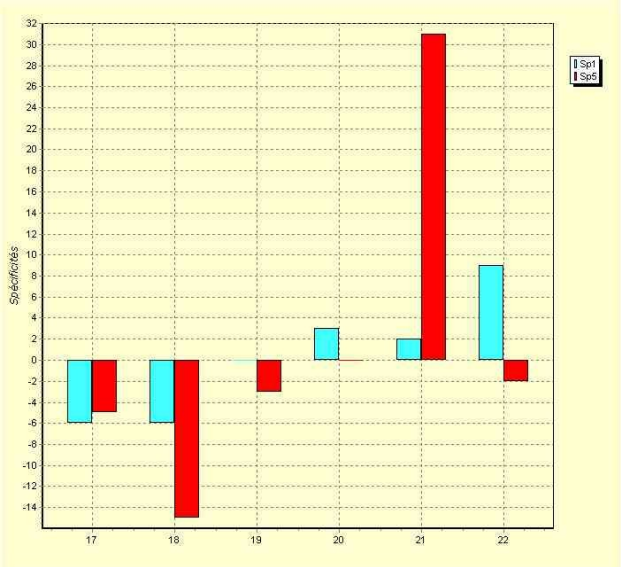
- du côté du *Oui* : *vous, gauche, Fabius, monde, tenants du non, extrême, Villiers, la France …*

- du côté du *Non* : *constitution, partisans du oui, je voterai non, médias, euros, démocratie, article(s) …*

Nous avons voulu savoir si ces lexiques spécifiques apparaissaient aux mêmes moments dans la chronologie du débat. À partir des vocabulaires dégagés comme caractéristiques pour les deux groupes de participants, nous avons construit deux unités textuelles que l'on appelle *types généralisés* (ou *tgen*) *spécifiques*[1], dont nous avons calculé la ventilation dans les six semaines qui composent le corpus : Le *tgen SP-oui* rassemble toutes les occurrences du corpus qui correspondent à une forme spécifique pour le groupe de participants qui se prononce pour le *oui*. Le tgen *SP-non* rassemble celles des occurrences qui correspondent au contraire à des formes spécifiques pour le groupe des participants qui se prononce pour le *non*[2].
On observe alors que chacun des deux camps connaît un moment d'expression plus marquée des termes qui font sa spécificité (la semaine 18 pour le camp du *non*, la semaine 21 pour celui du *oui*). Autre fait remarquable, le camp du *non* délaisse dans cette même semaine 21 le vocabulaire spécifique mis en place lors de la semaine 18. Ces constats sont confirmés par une nouvelle partition du corpus en douze parties (6 semaines x 2 opinions), c'est-à-dire que chaque partie rassemble dorénavant des messages produits au cours d'une même semaine par des internautes d'opinions identiques.
Ces calculs reposent sur les spécificités du « oui » et du « non » prises globalement. Or, nous avons dégagé, plus haut, quatre classes distinctives du « oui » et du « non ». Nous avons donc cherché la distribution chronologique des *Tgens* de ces classes d'arguments, que rapportent les graphiques ci-dessous.



Distribution chronologique des formes spécifiques des classes 3 et 6 (« non »)



Distribution chronologique des formes spécifiques des classes 1 et 5 (« oui »)

Il apparaît nettement que les deux classes d'arguments corrélées avec le « non » suivent une évolution comparable et conforme à la distribution globale des arguments du « non ». C'est un peu différent pour le « oui » : La classe 1 (des valeurs européennes) se trouve bien spécifique du « oui », et bien en fin de débat, mais de façon moins marquée que la classe 6 (des acteurs de la politique française), qui connaît une spécificité considérable la semaine même du scrutin.
Nos analyses montrent que le débat à propos du TCE ne se réduit pas à des formules simplistes selon lesquelles le « oui » serait européen et le « non » serait national.
Ainsi, les arguments du « non » semblent s'organiser sur le traité et font volontiers référence au texte lui-même, qui fait l'objet d'interprétations idéologiques, soit sur l'organisation des institutions

[1] Le Tgen, « ensemble d'occurrences sélectionnées parmi les occurrences du texte », généralise la notion de type unité dont on peut recenser les occurrences dans le texte. Sur les types généralisés, on consultera [Lamalle & Salem 2002].
[2] Nous avons sélectionné, pour cette expérience les unités qui ont un indice de spécificité supérieur à 4 (i.e. celles pour lesquelles la méthode des spécificités renvoie à une probabilité inférieure à 1/10000^{e}).

européennes, soit sur ses conséquences sociales et économiques (chômage, services publics, etc.). Par ailleurs, ces arguments s'installent très tôt dans le débat et ont tendance à faiblir par la suite.

Les arguments du « oui » reposent davantage sur une identification aux valeurs historiques fondatrices de l'union européenne et un recours aux symboles qui unissent les peuples et les citoyens. Le résultat le plus inattendu est certainement que c'est également du côté du « oui » que l'on trouve les références à la politique française et ses acteurs. Mais on observe également que ces arguments apparaissent très tard dans le débat : les partisans du « Oui » peinent à installer leur argumentation en début de confrontation et ce n'est que dans la dernière semaine qu'elle prend toute sa dimension.

Le recours à la statistique distributionnelle nous a permis de « faire parler » un corpus textuel, dans une acception que nous voudrions sans doute rattacher à une rhétorique plus psychanalytique que policière… Il n'en demeure pas moins que le rapport des sciences humaines et sociales à la matérialité textuelle, focalisé sur la dimension « verticale », se distingue nettement de l'approche linguistique ou littéraire, plus préoccupée de considérations « horizontales ». S'agissant de discours persuasifs, si ce n'est propagandistes, il ne s'agit pas de « comprendre » un corpus dans ses dimensions intra voire même intertextuelles, de rechercher une structure, d'identifier une esthétique, d'attribuer un genre ou de questionner l'une ou l'autre de ces notions. Le corpus en sciences humaines et sociales n'est qu'un accès possible à son contexte sociohistorique de production et n'a de réelle valeur qu'en tant qu'il nous renseigne sur ce contexte. Quant à l'analyse automatique, elle n'a d'intérêt qu'en tant qu'elle peut mettre en évidence des effets de sens qui, pour des raisons multiples, n'émergent pas d'autres analyses.

BIBLIOGRAPHIE

Le débat, n°136 : La France et le choc du 29 mai, septembre-octobre 2005.

LEBART, L., SALEM, A. 1994. *Statistique textuelle*, Paris, Dunod.

Les cahiers du CEVIPOF, n°42 : Le référendum de ratification du Traité constitutionnel européen : comprendre le « Non » français, juillet 2005.

MANGE, J., MARCHAND, P. à paraître. Oui ou non à la Constitution européenne : l'éloquence du forum, *Mots*.

MANGE, J., MARCHAND, P., SALEM, A. 2006. Débats sur la toile. Actes des Huitièmes Journées Internationales d'Analyse des Données Textuelles (Besançon, 19-21 avril). Presses Universitaires de Franche-Comté, coll. Les cahiers de la MSH Ledoux, série « archive, bases, corpus », pp. 667-676.

MARCHAND, P. 1998. L'Analyse du Discours Assistée par Ordinateur. Concepts, méthodes, outils. Paris, Armand Colin, coll. U, série psychologie.

MARCHAND, P. 2004. Psychologie sociale des médias. Presses Universitaires de Rennes.

MARCHAND, P. 2005. Le grand oral de Dominique de Villepin, Bulletin de Méthodologie Sociologique, n°87, pp. 80-85.

MARCHAND, P. à paraître. Concepts, méthodes, outils, in C. Gauzente & D. Peyrat-Guillard (éds.), *Analyse et fouille de données textuelles : applications en gestion*, Editions Management et Société, coll. « questions de société ».

REINERT, M. 1990. « ALCESTE, une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval », Bulletin de Méthodologie Sociologique, N°26.

SALEM, A. 1993. *Méthodes de la statistique textuelle*, Thèse, Université de la Sorbonne Nouvelle, Paris 3.

CONCORDANCES ET CONCORDANCIERS DE L'ART DU BON KWAC

Bénédicte PINCEMIN
CNRS / Université de Paris 13, LLI

SOMMAIRE

Résumé : *Les concordances sont un mode de présentation d'extraits de texte, contenant tous le même mot ou le même motif linguistique. C'est une méthodologie d'analyse textuelle séculaire, l'exemple même du "travail de bénédictin" avant le recours possible aux ordinateurs. Les concordanciers sont des outils informatiques produisant les concordances souhaitées à partir d'un corpus numérique. Le sujet qui nous intéresse est de comprendre cette remarquable pérennité de la concordance, et de repérer et analyser les transformations discrètes mais décisives apportées par leur calcul automatisé. Ce parcours conduit également à s'interroger sur une exploitation plus systématique et pertinente des possibilités ouvertes par l'ordinateur : quelle généralisation opératoire peut-on définir de la méthode des concordances ?*

Classiquement, dans les logiciels d'analyse textuelle, un calcul de concordance se définit par la détermination de trois paramètres : la donnée d'un pivot, à savoir le mot ou motif linguistique dont on veut étudier les occurrences en contexte ; la taille du contexte à visualiser ; et un critère de tri (éventuellement multiple) fixant l'ordre de présentation des contextes. Notre généralisation est spécifiante : plutôt que de multiplier les paramétrages et réglages, nous proposons de focaliser les ajustements sur ce qui fait la force des concordances calculées. Nous retenons donc pour une "bonne" concordance deux caractéristiques essentielles : les contextes présentés sur une ligne et superposés, de sorte à créer des effets visuels d'alignement vertical, d'une part ; et le tri des lignes de contexte, sur un ou plusieurs éléments, d'autre part.

Chemin faisant, nous écartons donc clairement d'autres relevés d'occurrences quelquefois appelés concordances. En nous inspirant très librement de désignations de types d'index en sciences de l'information, nous distinguons le KWOC (keyword out of context), liste d'attestations ; le KWAC (keyword and context), concordance telle que nous l'entendons, avec son dispositif de regroupement visuel par superposition, alignement vertical et tri ; le KWIC (keyword in context), relevé de contextes avec un choix plus libre de la taille de ceux-ci ; le KWUT (keyword up to text), où les occurrences sont repérées au fil du texte. Ces quatre modes de relevé d'occurrences sont complémentaires et gagnent à être cultivés et utilisés pour leurs spécificités.

La concordance "papier" traditionnelle et les sorties d'un concordancier diffèrent davantage que par leur mode de production ; mieux, chacune tire le meilleur parti des spécificités de leur processus de construction : travail de synthèse et guide interprétatif pour la première, régularité, polyvalence et dynamique pour la seconde. Cependant, par des voies différentes, concordances manuelles et concordances calculées servent le même principe herméneutique fondamental : la mise en évidence des parallélismes et des contrastes dans les contextes de l'item étudié.

Dans cette optique, pour la concordance (KWAC), le réglage de la taille des contextes devient secondaire (voire nuisible, car dénaturant potentiellement la concordance ; c'est plutôt une caractéristique du KWIC). À l'inverse, le mode de définition du pivot gagne à être affiné, par l'introduction d'une décomposition en zones, qui permettent de démultiplier et d'assouplir les dispositifs d'alignement et de rapprochement visuels (Pincemin et al. 2006).

1. Comprendre le succès des concordances

Il est frappant d'observer la popularité toujours actuelle des concordances, telle que manifestée par l'abondance des concordanciers[1] et l'omniprésence de la fonction de calcul de concordances dans les logiciels d'analyse de texte. Pourtant, la procédure informatique est simple, relativement à d'autres techniques d'analyse textuelle : il s'agit d'une réorganisation du matériau textuel par une indexation machinale (relevé systématique des mots[2]), un tri formel (alphabétique), et une présentation astucieuse. L'état de l'art des outils d'analyse textuelle révèle et démontre toute une gamme de procédures plus élaborées, tout particulièrement dans le domaine des statistiques textuelles (lexicométrie puis textométrie)[3]. Mais ces techniques sont encore peu diffusées, restent surtout l'apanage de logiciels de recherche français (tels que Weblex, Hyperbase, Lexico). Les concordances restent de fait l'outil de référence hors de la communauté des statistiques textuelles, comme au plan international, pour les outils de consultation et d'analyse de corpus numériques.

Ce succès témoigne à la fois de la complémentarité de cette technique relativement aux développements innovants de la statistique textuelle, et, dans l'absolu, du bien-fondé herméneutique, de la pertinence, de l'efficacité des concordances pour l'étude des textes. D'où l'intérêt d'examiner de plus près leur conception et leur fonctionnement : sur quels principes les concordances sont-elles fondées ? Quelles sont les éventuelles variantes de réalisation ? Y a-t-il des versions plus intéressantes que d'autres, pour quelles raisons ou/et sur quels plans ?

2. Des trois paramètres fondamentaux des concordanciers à une définition des concordances

L'expérience du recours à différents logiciels permet d'abord de repérer une définition concrète de la concordance, à travers les paramétrages prévus. En pratique, une concordance se calcule à partir de trois éléments :

(i) un pivot, typiquement un mot (mais cela peut être généralisé à toutes sortes d'items dont les occurrences sont repérables par le logiciel considéré : lemme, expression, etc.), aux contextes duquel on s'intéresse ;

(ii) une taille de contexte (éventuellement détaillée en contexte gauche et contexte droit, qui respectivement précède et suit le pivot), typiquement la longueur d'une ligne pour le mode d'affichage utilisé ;

(iii) un ordre de présentation des contextes sélectionnés, typiquement ordre de présence dans le corpus, ou tri alphabétique sur le mot qui précède le pivot (tri gauche) ou sur celui qui le suit (tri droit).

Cette première généralisation rend ensuite sensible aux implémentations qui restent en-deçà de ces possibilités, ou à l'inverse qui étendent la portée d'un paramètre. Dans le premier cas (implémentation pauvre), l'incidence de l'occultation d'un paramètre n'est pas forcément négative, si elle simplifie l'application informatique en la centrant d'emblée sur des valeurs de paramètre suffisantes et efficaces pour les domaines d'usage prévus. Par exemple, pour l'étude linguistique du lexique telle que la pratique notre laboratoire, selon une approche de type sémantique distributionnelle (le sens d'un mot est caractérisé par ses contextes d'emploi, notamment par ses dépendances syntaxiques), l'absence du second paramètre (si le logiciel affiche toujours des contextes d'une ligne, de longueur fixe, optimale) s'avère souvent moins pénalisante que l'impossibilité de trier les contextes, car les alternatives permises par ce troisième paramètre répondent à une plus grande diversité de questionnements.

De fait, la définition par les paramètres laisse implicite une caractéristique essentielle des concordances : l'alignement vertical, sur une colonne (généralement centrée), des occurrences du pivot. Cet alignement en colonne, associé au tri des lignes de contexte permettant de rapprocher les contextes analogues, souligne visuellement, par superposition et répétitions, les convergences et les divergences de formulation. Cette présentation s'avère un outil heuristique extrêmement efficace pour la lecture des contextes proches et l'observation globale des constructions dans lesquelles s'insère le mot étudié.

---

[1] Par exemple : AntConc, KWICFinder, MonoConc, TACT…

[2] La linguistique montre toute la complexité de la notion de "mot" ; on s'en tient ici aux définitions opératoires, intuitives et simplifiées, qui sont utilisées en statistique textuelle, en termes de chaînes de caractères non délimiteurs maximales (Lebart & Salem 1994).

[3] Sans même entrer ici dans le domaine, très développé et actif, des analyses faisant appel à des modélisations linguistiques ou/et sémantiques.

Ainsi, alors que l'usage désigne par KWIC (KeyWord In Context) la technique des concordanciers, nous préférerions parler ici de KWAC (KeyWord And Context), pour souligner que les relevés de contextes sont organisés autour du mot étudié et à partir de lui. La concordance est concordance parce qu'elle est visuellement ancrée sur le pivot qui structure le parcours de lecture. La disposition invite à faire du mot-clé (*KeyWord*) le point d'entrée ou du moins un repère central constant, et (*And*) son contexte (*Context*) immédiat est disposé de part et d'autre. Ce faisant, la désignation de KWIC reste disponible pour distinguer une forme complémentaire de relevé d'occurrences, pour laquelle le mot (*KeyWord*) reste davantage inséré (*In*) dans son contexte (*Context*), la présentation lui donnant un rôle moins dominant (cf. § 3).
La pratique des concordanciers rend enfin sensible à l'importance interprétative de l'indication, pour chaque contexte extrait, d'une référence intelligible permettant de le situer au sein du corpus. En effet, l'interprétation se nourrit d'informations locales (les mots dans le voisinage étudié) mais aussi globales : par exemple, de quel texte est tiré ce contexte ? à quel moment a-t-il été écrit, de quelle source provient-il ? Autant d'indications sans lesquelles l'interprétation, artificiellement privée d'une grande part du contexte, est considérablement appauvrie. Les concordanciers peuvent maintenant recourir utilement à l'hypertexte pour donner directement accès au point du texte dont provient le contexte. Ceci ne périme cependant pas une mention mnémonique, choisie pour apporter les informations intertextuelles jugées utiles, et que l'on a sous le regard en même temps que l'on balaye les occurrences en contexte. Ces références de localisation se contextualisent en outre mutuellement : on y lit le passage d'une partie du corpus à une autre, la richesse ou la pénurie d'attestations selon les localisations.
Ces observations permettent de formuler une définition synthétique du (bon) KWAC : un corpus étant fixé, une concordance (ici KWAC) est la liste de toutes les occurrences d'un pivot, alignées verticalement en colonne, entourées de part et d'autre par leur contexte, accompagnées d'une référence indiquant de façon pertinente leur positionnement dans le corpus, et triées selon un critère pertinent pour l'analyse.

3. KWOC, KWAC, KWIC et KWUT : une typologie des relevés d'occurrences

Plus généralement, une certaine confusion règne entre différentes formes de relevés lexicaux au sein d'un texte numérique. En nous inspirant très librement d'une formulation mnémonique ayant cours dans les sciences de l'information (cf. § 5), nous proposons de différencier les KWOC, les KWAC, les KWIC et les KWUT.
Les KWOC sont les KeyWord Out of Context. Il s'agit de la liste des différentes attestations correspondant à la requête soumise par l'utilisateur, qui fait office de filtre sur le vocabulaire complet. Comme il n'y a pas de contexte (sinon celui, interne, de l'occurrence elle-même, qui peut être une expression composée), les KWOC ne se laissent guère confondre avec les concordances et autres relevés de contexte, même s'il y a une certaine parenté théorique et technique entre ces procédures. La force du KWOC vient de son caractère synthétique. L'effacement du contexte permet une réduction supplémentaire à celle d'une concordance à taille de contexte nulle : c'est le regroupement des différentes occurrences ayant la même forme. Pour un même choix de pivot, un KWOC est donc d'une longueur bien moindre que celle d'une concordance, en évitant certaines redondances si l'attention se porte moins sur le contexte que sur le pivot. En chiffrant les répétitions condensées sous une même forme et en offrant des tris tant sur les formes (alphabétique, canonique) que sur les fréquences (tri dit hiérarchique), les relevés KWOC sont très efficaces pour jauger le relief quantitatif de différentes formulations, pics attracteurs comme lacunes surprenantes et significatives (l'équipe Hubert de Phalèse a même trouvé un nom à ces dernières (fréquences nulles) : les *nullax*, de la même manière qu'on a le cas particulier remarquable des hapax, occurrences de fréquence 1). Enfin, remarquons que ce "hors contexte" n'est en rien décontextualisé, le contexte immédiat ne s'éclipse que pour mieux manifester le contexte textuel ou intertextuel, grâce aux références de localisation, qui résument le profil de répartition des occurrences. Malgré son nom, le relevé KWOC est pleinement lié au corpus, dont il pointe efficacement des caractéristiques expressives profondes et pas toujours perceptibles.
Les KWAC sont les KeyWord And Context. Les occurrences du pivot étudié sont présentées avec leur contexte et alignées verticalement pour former une colonne, de telle sorte qu'une double lecture est possible, verticale (liste des formes du pivot : KeyWord) et (And) horizontale (contextes : Context). Cette superposition est essentielle car elle met en évidence les répétitions plus ou moins massives et les variantes, tout particulièrement lorsqu'elle est associée à un tri alphabétique

des contextes. Elle donne appui à un parcours de lecture des voisinages d'un mot complètement nouveau et très riche, mettant très efficacement en lumière des régularités d'usage pas toujours perceptibles dans une lecture linéaire classique. C'est aux relevés présentant une telle heuristique visuelle de lecture, par "empilement", que nous réservons l'appellation de concordance. Comme nous l'avons vu, selon cette perspective, la définition pratique des concordances par les trois paramètres -pivot / taille du contexte / ordre de présentation- doit être révisée. S'ajoute une caractéristique définitoire : l'alignement vertical des occurrences du pivot. La définition en est alors corrigée : l'effet d'empilement ne se réalise pleinement que si les contextes se présentent chacun sur une ligne, la taille du contexte n'a alors plus vraiment lieu de rester un paramètre.

Les KWIC sont les KeyWord In Context[1]. Il s'agit d'un relevé des contextes d'un mot, en privilégiant souvent des contextes assez larges, de l'ordre de plusieurs lignes ou phrases, ou d'un paragraphe par exemple. Le mot est mis en valeur afin de donner appui à cette résonance interprétative entre le mot (KeyWord) et son contexte (Context). Une telle taille de contexte, élargie et assouplie, annule en grande partie les effets de lecture permis par un alignement vertical sur le pivot et une superposition des contextes, qui reste l'apanage du KWAC. On retrouve ainsi les deux premiers paramètres des concordances (choix du pivot et de la taille des extraits), mais le troisième n'est plus toujours pertinent. La force du KWIC, c'est l'optimisation potentielle des contextes locaux d'observation : toute liberté est donnée sur la taille ou la nature des contextes, et ceux-ci peuvent être restitués selon leur présentation usuelle (en préservant par exemple ainsi l'information visuelle de l'occurrence en début / fin de paragraphe). Lorsqu'un même contexte contient plusieurs occurrences, il n'y a évidemment pas lieu de le répéter à l'identique. Les concentrations locales d'occurrences (répétition du pivot) sont donc particulièrement bien rendues (visualisation de la disposition dans le contexte local), sans alourdir inutilement le relevé. Un "bon" KWIC a une très forte valeur interprétative : il s'agit d'un recueil de contextes délimités de façon pertinente, de véritables unités sémantiques pour l'interprétation des attestations, bénéficiant d'une certaine autonomie. Chaque contexte peut devenir comme un petit texte : ce n'est pas sans rappeler la pratique des recueils de citations, ou des morceaux choisis. Sans renoncer au contexte du corpus (normalement motivé, par construction), l'interprétation peut s'enrichir en jouant de focalisations à géométrie variable, à la fois souples et contrôlées (tout découpage n'est pas pertinent).

C'est entre les KWIC et les KWAC que les confusions sont les plus répandues : le KWIC pourrait n'être vu que comme un KWAC au contexte plus long ; et, dans les concordanciers usuels, le réglage de la taille de contexte pourrait être un mode de basculement insensible du KWAC au KWIC. En fait, KWIC et KWAC sont deux spécialisations complémentaires du concordancier, donnant chacune toute sa mesure à l'un des paramètres, sans que l'un bride ou affaiblisse l'autre : le KWIC libère l'éventail des tailles et types de contextes, et le KWAC tire le meilleur parti des effets visuels, heuristiques et herméneutiques, offerts par les tris. Les contextes sur une ligne et superposés du KWAC se prêtent bien à une organisation selon une progression continue, et les contextes plus textuels du KWIC se structurent assez naturellement en classes, par regroupements, rapprochements et contrastes.

À la différence du KWAC qui est par nature centré sur le voisinage immédiat du pivot, le KWIC permet d'explorer un contexte focalisé tout en étant délimité plus souplement. Le KWIC peut ainsi être une réponse bien adaptée à certains questionnements sémantiques par exemple, lorsque les incidences sémantiques recherchées sont diffuses et peu cadrées, alors que le KWAC pourra être plus efficace pour l'étude de relations ancrées dans le lexique ou la syntaxe, à portée plus limitée et aux formes plus régulières, mieux saisissables par des tris. Un cas intéressant est celui de la technique de concordance enrichie mise au point par Evelyne Bourion (2001). Dans une perspective d'analyse sémantique, les mots statistiquement significativement associés au pivot (les corrélats) sont mis en valeur typographiquement ; les lignes de concordance sont triées en fonction de la densité et de la force de ces associations sémantiques potentielles calculées. Or ces corrélats ne créent que peu ou pas d'effets d'alignement visuels, leur relation, sémantique, est souple en terme de construction syntaxique comme de distance au pivot. Ces relevés sont donc un cas limite de KWAC : l'effet de superposition est affaibli (pas d'empilement en colonne des

[1] KWIC est une formule plus courante que les deux autres (KWOC et KWAC -KWUT est un néologisme de notre cru), et est souvent utilisée pour désigner les concordances. Nous nous écartons donc de l'usage, pour tirer bénéfice de la typologie soulignée par la formulation mnémonique quadruple KWIC KWAC KWOC KWUT.

corrélats), cependant présent (si le contexte est sur une ligne et grâce à la mise en valeur typographique qui attire le regard et facilite l'établissement de correspondances).
Les KWUT sont les KeyWord Up to Text. Le mot étudié est mis en valeur au sein de son contexte, au fil du texte. Autrement dit, le texte est affiché dans son entier, dans son déroulement intégral, et les occurrences sont repérées dans ce contexte global. Le KWUT donne accès à l'organisation des occurrences à l'échelle textuelle, à leur disposition dans la linéarité textuelle. Il peut par exemple y avoir des effets de regroupement dans une zone du texte, positionnée à tel endroit (début, fin du texte,...) ou à l'inverse d'évitement de telle ou telle partie. Comme on revient au texte dans son intégralité avec sa mise en forme, a minima le marquage de ses grandes subdivisions, le KWUT renseigne aussi sur la saillance éventuelle de telle ou telle occurrence du fait de sa position, notamment aux frontières : par exemple mot dans la phrase qui débute un chapitre, ou "mot de la fin". Pour être exploitable, la présentation KWUT devrait être couplée à un dispositif de repérage rapide des occurrences, de sorte à ne pas être contraint de parcourir linéairement tout le texte pour visualiser toutes les occurrences. Généralement on dispose d'une fonctionnalité de saut d'une occurrence à l'occurrence suivante (ou précédente). Une autre possibilité, complémentaire car plus textuelle, mais encore peu répandue, est la présentation conjointe (et hypertextuellement liée) d'une vue d'ensemble de la distribution des occurrences du pivot dans le texte. La "carte des paragraphes" du logiciel Lexico 3, ou l'histogramme marginal proposé par Pincemin[1], sont des exemples de représentations textuelles visuelles et synthétiques de ce type.
À ce parcours du KWOC au KWUT, s'associe une extension progressive des contextes : lexie et syntagme pour le KWOC, syntagme à période avec le KWAC, période ou paragraphe avec le KWIC, texte voire intertexte dans le KWUT. L'intertexte ne semble pas motiver une forme de relevé en contexte supplémentaire, mais il s'introduit de façon plus ou moins saillante dans les autres relevés. Un KWOC peut détailler la répartition des formes ou/et des fréquences dans les différentes parties du corpus (textes, catégories), et donner ainsi une vision globale des attestations par rapport à l'articulation intertextuelle du corpus. Si le corpus présente une organisation orientée, les KWAC sont quant à eux bien appropriés pour rendre compte synthétiquement des effets de succession et d'évolution, grâce au tri des contextes selon l'ordre du corpus. Les contextes KWIC se prêtent particulièrement bien à une structuration par regroupements : les divisions intertextuelles peuvent éclairer l'analyse des relevés KWIC, et réciproquement les contextes d'attestation KWIC peuvent caractériser les articulations du corpus quant à l'usage de telle ou telle forme.

4. Illustration : propositions pour les concordanciers sur corpus multilingues parallèles alignés

Considérons le cas où l'on dispose d'un corpus numérique formé de différentes versions d'un même texte, traduit en plusieurs langues. Les correspondances d'une langue à l'autre sont accessibles entre passages plus ou moins fins : selon la nature du texte et la procédure d'alignement cela peut être de l'ordre du paragraphe, de la phrase ou du mot.
Le KWOC peut être utile par exemple si l'on dispose d'une forme d'alignement au niveau des mots, pour lister synthétiquement les correspondances lexicales d'un mot dans une autre langue dans le contexte du corpus étudié.
Le KWAC est très efficace pour étudier les usages d'un mot à l'intérieur d'une langue, notamment les constructions grammaticales dans lesquelles il s'inscrit, et ses associations sémantiques proches (comme les choix des qualificatifs).
Pour l'étude multilingue et simultanée, parallèle, des contextes d'un mot, un KWIC semble plus pertinent qu'un KWAC, car l'effet de superposition et de tri est très affaibli et souvent non pertinent lorsque les langues sont mêlées ou intercalées. Une présentation efficace serait donc un KWIC sur plusieurs colonnes (une par langue) relativement étroites, de sorte à donner de l'épaisseur visuelle aux contextes, et faciliter ainsi l'observation des correspondances et des décalages de positionnement, eux-même révélateurs des variantes de formulation.
Autrement dit, un concordancier multilingue n'est pas (ne devrait pas être) la simple application d'un concordancier usuel à un corpus multilingue. Mais aussi, une concordance sur corpus aligné ne gagne rien à être un alignement de concordances. Même en contexte multilingue, la

[1] Voir par exemple Pincemin B. (2001). « Résoudre la surcharge informationnelle sans décontextualiser », in Stéphane Chaudiron et Christian Fluhr (éds), 3ème colloque du chapitre français de l'ISKO « Filtrage et résumé automatique de l'information sur les réseaux », Université Paris X, 5-6 juillet 2001, pp. 149-158.

concordance se centre et se calcule sur une langue. Cette affirmation n'exclut pas bien sûr la possibilité de consulter facilement, depuis la concordance, tel contexte aligné dans telle autre langue ; elle n'exclut pas non plus la possibilité de calculer, en se basant sur l'analyse d'une concordance dans une langue, une autre concordance dans une autre langue, avec des facilités pour les visualiser conjointement et naviguer de l'une à l'autre.

5. Retour épistémologique et terminologique sur les KWIC

Notre proposition de typologie des relevés d'attestations en KWOC, KWAC, KWIC et KWUT a l'avantage d'être mnémonique, tout en contrastant bien différents modes de présentation complémentaires. Mais elle a l'inconvénient d'aller à contre-courant de l'usage particulièrement bien établi et largement diffusé pour les KWIC, et des notions apparentées de KWOC et de KWAC, classiques bien connus des experts en gestion de l'information. Il nous semble nécessaire de rendre compte ici de ce contexte épistémologique.

Le concept de KWIC, et sa désignation, ont été définis par Hans Peter Luhn, à la fin des années 50. Il s'agissait de construire un index des titres de publications scientifiques, selon une procédure simple automatisable (afin de disposer d'un index très complet actualisé très fréquemment). Les mots-clés sont alors les mots lexicaux des titres (on écarte simplement les mots grammaticaux et éventuellement des mots généraux jugés ici peu significatifs). Ces mots sont alignés en colonne au centre de la page, entourés de part et d'autre par leur contexte dans chaque titre. Ces relevés sont triés selon les mots-clés centrés, de façon à former l'index : les différents mots-clés apparaissent alors dans l'ordre alphabétique, avec pour chacun le relevé des différents titres dans lequel il apparaît. Voici une illustration de ce procédé appliqué à deux intitulés de colloque :

SdN06	Semaine du	**Document**	numérique
Albi06	Colloque	**Documents**	numériques et interprétation
Albi06	Colloque Documents numériques et	**interprétation**	
SdN06	Semaine du Document	**numérique**	
Albi06	Colloque Documents	**numériques**	et interprétation

Fig. 1 : Exemple de KWIC (au sens traditionnel)

Le procédé consiste donc à faire tourner astucieusement les chaînes de caractères des titres ("rotated strings"). Le KWOC et le KWAC sont alors des variantes à partir de ce principe[1].

Le KWAC (KeyWord Alongside Context, ou KeyWord And Context[2]) a pu être imaginé pour optimiser le volume. En effet, l'index KWIC sur les titres (contextes courts cités dans leur entier) n'occupe en moyenne qu'une moitié de chaque ligne[3]. Pour éviter ces blancs liés au centrage du mot-clé, le mot-clé peut être placé en tête de ligne, suivi de son contexte ultérieur. Puis, après un séparateur conventionnel (éventuellement tout simplement le point), le contexte antérieur est

---

[1] D'après les documents que j'ai pu consulter (en particulier Hjørland B. (2006). KWIC / KWAC / KWOC. http://www.db.dk/bh/lifeboat_ko/SPECIFIC%20SYSTEMS/kwic_kwac_kwoc.htm (28/01/2006)
qui renvoie lui-même à Buckland M. (2003). Organization of Information in Collections : Verbal Access. http://www.sims.berkeley.edu/courses/is245/s03/verbal.html
et Taylor A. (2000). Wynar's Introduction to Cataloging and Classification. 9th edition. Littleton, colo.: Librairies Unlimited. pp. 408-411),
Luhn a inventé le KWIC vers 1958 (Sékhraoui 1995 cite par exemple : Luhn H.-P. (1959). "Keyword-in-context index for technical literature (KWIC index)", American Documentation, 11 (4), pp. 288-295. Reprinted in Hays, David G. (ed.), Reading in Automated Language Processing, New York, 1966, pp. 159-167).
Le KWOC et le KWAC auraient été imaginés ultérieurement.

[2] (Salton & McGill 1983) ne consacrent que quelques lignes à ce sujet. C'est la première présentation (et longtemps la seule) que j'ai eue des KWIC, KWAC et KWOC, elle a ainsi influencé ma réflexion, mais est peut-être trompeuse. En effet, le KWAC y est explicité comme KeyWord And Context, alors que les références données en note précédente parlent de KeyWord Alongside Context. Le tableau illustratif donné en exemple ne concerne qu'un KWIC et un KWAC, mais ce KWAC est un KWOC au sens des références précédentes. Le KWOC n'est pas décrit ni illustré.
D'une manière générale, le KWAC est beaucoup moins connu que le KWOC, lui-même plus marginal que le KWIC.

[3] La ligne KWIC serait occupée un peu plus qu'à moitié si tous les titres étaient de même longueur et si cette longueur correspondait à la largeur de la page. En pratique, comme ces conditions sont *a priori* assez loin d'être réalisées, la ligne KWIC est plus qu'à moitié vide.

donné. Cette tactique traduit très directement cette idée de « chaînes tournantes » (rotated strings). Outre le gain de place, l'association entre le mot-clé et sa référence de localisation est facilitée. Voici une illustration du KWAC sur les mêmes données que le KWIC précédent :

SdN06	**Document** numérique. Semaine du
Albi06	**Documents** numériques et interprétation. Colloque
Albi06	**interprétation** . Colloque Documents numériques et
SdN06	**numérique** . Semaine du Document
Albi06	**numériques** et interprétation. Colloque Documents

Fig. 2 : Exemple de KWAC (au sens traditionnel)

Le KWOC aurait alors pu être introduit pour garder le principe d'optimisation spatiale du KWAC tout en ménageant un meilleur confort de lecture, en évitant la coupure abrupte des titres. Le mot-clé est tout simplement repris en tête de chaque ligne (comme le KWAC, pour éviter le centrage de type KWIC), suivi du titre dans son entier. Le mot-clé est donc bien comme sorti de son contexte pour être répété en tête de ligne. Une deuxième forme de KWOC factorise le mot-clé en le sortant encore davantage des lignes de contexte : si les mots-clés sont en général présents dans plusieurs titres, on peut trouver plus claire et plus concise une présentation où le mot-clé n'est donné qu'une fois, en introduction au groupe de titres concernés. On réinvente alors la présentation d'un index traditionnel.

SdN06	**Document**	Semaine du Document numérique
Albi06	**Documents**	Colloque Documents numériques et interprétation
Albi06	**interprétation**	Colloque Documents numériques et interprétation
SdN06	**numérique**	Semaine du Document numérique
Albi06	**numériques**	Colloque Documents numériques et interprétation

Fig. 3 : Exemple de KWOC (au sens traditionnel), première forme (sans factorisation)

Document(s)
Semaine du Document numérique *(SdN06)*
Colloque Documents numériques et interprétation *(Albi06)*
Interprétation
Colloque Documents numériques et interprétation *(Albi06)*
Numérique(s)
Semaine du Document numérique *(SdN06)*
Colloque Documents numériques et interprétation *(Albi06)*

Fig.4 : Exemple de KWOC (au sens traditionnel), seconde forme (avec factorisation[1])

Enfin, le Double KWIC (proposé par Anthony E. Petrarca et W. Michael Lay au début des années 1970, et dont la désignation est peu connue) est à l'origine de ce perfectionnement décisif du KWIC : le tri des lignes non seulement en fonction du pivot (c'est le premier tri), mais aussi en fonction de son contexte. C'est en effet un deuxième tri qui permet les effets de superposition de contexte si utiles. Actuellement, bien des concordanciers se présentent comme des générateurs de KWIC, et en proposent, sans doute sans le savoir, une forme avancée (issue du principe du Double KWIC) qui s'est imposée par sa pertinence.

6. Originalité et apports de la pratique séculaire des concordances

Après son parcours au sein des concordanciers de toutes origines et de toutes formes, le chercheur venant de l'informatique textuelle peut encore découvrir du nouveau en se penchant sur les pratiques de concordances "d'avant l'informatique". Comme dans bien d'autres cas[2], l'identité

[1] Pour mettre en évidence la factorisation à l'échelle de notre exemple, les mots-clés ont été lemmatisés, de sorte à pouvoir grouper sous une même entrée le singulier et le pluriel d'un même mot. On peut bien sûr construire un KWOC avec factorisation et sans lemmatisation, c'est même le cas dans la lignée la plus directe des KWIC et KWAC, où les techniques peuvent être très simples et se limiter à des réorganisations de chaînes de caractères.

[2] Celui de l'indexation ou des mots-clés par exemple, l'indexation automatisée du texte intégral dans les corpus documentaires numériques produisant des mots-clés fondamentalement différents, dans leur nature

de désignation -les *concordances* bibliques des moines *vs* les *concordances* produites par un logiciel d'étude de la Bible- masque en fait des réalités sensiblement différentes[1].
Pointons particulièrement ici une différence qui touche à la définition que nous avons donnée des concordances : l'alignement vertical (sur une colonne) du mot pivot, avec la superposition des contextes ligne à ligne, qui est une caractéristique majeure de la concordance "KWAC", ne se retrouve pas dans la concordance traditionnelle. Faudrait-il alors plutôt reconnaître dans les concordances manuelle un ancêtre du KWIC[2] ? Et comment expliquer que l'idée de cette technique de mise en forme, si essentielle et si utile pour le KWAC, ait échappé à des générations d'érudits et d'experts de l'analyse des textes ? C'est que le parallélisme, induit visuellement par l'alignement et la superposition des contextes dans le KWAC, n'en est pas moins présent dans les concordances traditionnelles, où il se dessine d'une autre façon. En effet, les contextes d'occurrence d'un mot (pivot) sont généralement organisés par des regroupements sémantiques. Or ces regroupements sont bien souvent corrélés à des similitudes de contextes d'usage, notamment de construction, si bien que la lecture des contextes cités sous une rubrique donnée concentre et rassemble des voisinages récurrents. Les effets de parallélisme éventuels sont favorisés aussi par la brièveté voulue des contextes, en général sur une ligne (comme les KWAC).
Plus fondamentalement, le principe herméneutique à la base même des concordances, c'est bien celui de rapprochement des "passages parallèles". La concordance met en relation, dans un corpus à intertextualité dense, des parties de texte qui entrent en connivence par l'usage d'un même mot. L'interprétation se nourrit de cette considération simultanée de contextes ressemblants, et la concordance est un outil venant au renfort de la mémoire, qui déjà, consciemment et inconsciemment, tisse ces liens. D'ailleurs, le second grand type de document outil herméneutique qui fait le pendant à la concordance, c'est la synopse qui, comme son nom l'indique, présente côte à côte des passages textuels dispersés dans le corpus mais proches par leur sens et leur contenu. La synopse et la concordance sont deux points d'entrée d'une même pratique de rapprochements de contextes, la synopse partant du texte, la concordance du mot[3].
Les concordances "papier" traditionnelles et les sorties d'un concordancier diffèrent davantage que par leur mode de production ; mieux, chacune tire le meilleur parti des spécificités de leur processus de construction. La concordance construite manuellement dose le détail des relevés et la synthèse des informations (désambiguïsation et lemmatisation contrôlée, abstraction des constructions grammaticales, regroupements sémantiques, sélectivité...). Des contextes interprétatifs pertinents sont délimités à tous les niveaux (entre les entrées, à l'intérieur d'une entrée, choix des citations)[4]. Stable et globale (alors que le concordancier produit des vues dynamiques et locales), la concordance éditée implique des choix, dont la pertinence suppose une réflexion humaine (et non un traitement machinal). Œuvre d'auteur, elle communique une intelligence du texte, une lecture qui fait autorité. Le concordancier s'inscrit en complémentarité : le calcul assure un relevé systématique, régulier et exhaustif. Les tris et alignements verticaux suggèrent des regroupements des contextes par des parallélismes quelquefois inattendus et révélateurs. Grâce à la puissance et à la disponibilité du calcul, l'analyse textuelle se construit dynamiquement, à la croisée de multiples concordances, en variant les entrées et les tris : l'automatisation de la procédure ne dispense pas l'utilisateur d'une certaine habileté herméneutique, pour trouver un éventail de points de vue pertinents et éviter la dispersion. Plus fondamentalement, les concordances tirent parti de l'écriture en tant que technique d'analyse et support de réflexion, et les concordanciers ouvrent de nouvelles perspectives apparentées aux caractéristiques des traitements numériques[5]. Cependant, par des voies différentes, concordances

---

et leur fonctionnement, de ceux affectés à un texte lors d'un catalogage par un documentaliste.

[1] Dans cette partie nous résumons une étude qui pourra être développée dans une autre publication.

[2] KWIC : au sens que nous lui avons donné en § 3.

[3] D'autres formes de documents traditionnels trouvent un écho dans nos KWOC, KWAC, KWIC et KWUT. L'*index* s'apparente clairement au KWOC. Les *tables* recouvrent des réalités diverses, s'approchant généralement d'un KWIC avec un contexte minimal. Le *recueil de citations* est également un genre de KWIC, manuel, sélectif et non verbal (l'entrée qui définit un regroupement est une désignation thématique qu'on ne retrouve pas nécessairement littéralement dans les citations).

[4] Ces délimitations de contextes de tous ordres sont véritablement une difficulté pour un traitement automatique, or la contextualisation joue un rôle majeur pour l'analyse du sens d'un mot ou d'un texte.

[5] Nous évoquons ici le concept de *raison graphique* développé par Jack Goody, et son correspondant, la *raison computationnelle*, proposé et analysé par Bruno Bachimont.

manuelles et concordances calculées servent le même principe herméneutique fondamental : la mise en évidence des parallélismes et des contrastes dans les contextes de l'item étudié.

7. Une proposition technique d'amélioration des concordanciers : les zones

Pour renforcer encore les atouts spécifiques du concordancier (qui tiennent aux effets visuels d'alignement, de superposition et de répétition), nous avons proposé d'introduire la notion de "zones" dans la technique de construction des concordances. Nous résumons ici la présentation développée, commentée et illustrée dans (Pincemin 2006).

La sélection du pivot se détaille comme une séquence de zones successives : autrement dit, le pivot n'est pas un bloc, mais il est structuré, et composé d'une suite d'éléments individuellement identifiables et potentiellement actifs pour la construction de la présentation des résultats. L'intérêt de ce découpage en zones est de pouvoir indiquer, pour chacune d'elles, *(i)* si elle est le lieu d'un empilement (en formant une colonne), *(ii)* si son unité est soulignée par une mise en valeur typographique et laquelle (par exemple couleur, gras), *(iii)* si elle fait l'objet d'un tri et dans ce cas sur quelle dimension descriptive et de quel type (alphabétique, hiérarchique i.e. par fréquence décroissante, canonique i.e. suivant un ordre conventionnel). Les contextes gauche et droit disposent également d'une possibilité de tri. Au final le tri de la concordance se définit en fixant un ordre d'application des tris des zones et contextes concernés, s'il y en a effectivement plusieurs.

Un tel concordancier nouvelle génération reprend bien toutes les possibilités de tri proposées dans les concordanciers actuels. Il permet le tri sur des mots distants, tout en maîtrisant mieux la portée des tris. En particulier, le découpage du pivot en zones permet un repérage plus fin que la position en nombre de mots par rapport au pivot, puisqu'on peut par exemple s'appuyer sur l'étiquetage du corpus ou prendre en compte le contexte.

Les zones sont donc bien au service d'un bon KWAC : les effets d'alignement vertical sont démultipliés et mieux caractérisés.

8. Vers une compréhension linguistique de la puissance herméneutique des concordances

Le succès persistant des concordances témoigne de leur efficacité et de leur pertinence pratique. Et cette pertinence heuristique se comprend très bien au regard d'une sémantique textuelle et différentielle (Rastier 2001), selon laquelle le sens d'un mot, et plus généralement la construction d'une unité linguistique et son parcours interprétatif, se déterminent à partir de ses contextes de tous ordres, de leur rapprochement et de leurs contrastes. La concordance se présente en effet comme un instrument privilégié de l'étude de multiples contextes :

(i) le contexte syntagmatique, bien sûr, par le voisinage immédiat de chaque attestation du mot étudié ;

(ii) le contexte paradigmatique, par les liens possibles aux diverses éditions, aux traductions et à la formulation originale, par le choix des entrées et l'indication éventuelle de renvois, également par le voisinage ou le regroupement des entrées ;

(iii) les parallélismes synoptiques, par le rapprochement visuel des autres contextes du même mot, éventuellement accentué par des alignements verticaux (superposition en colonne sur le mot commun) et des tris (rapprochement et superposition des contextes identiques et mise en évidence des points de divergence) ;

(iv) le contexte textuel et intertextuel, par la référence qui localise l'extrait cité dans le corpus, et par la perception quantitative de la répartition des attestations dans le corpus.

Notre parcours d'analyse a voulu mettre en évidence des observations ainsi pertinentes pour la conception de concordances. Insistant sur le danger de généralisations confuses, notre cheminement invite à cultiver les atouts propres à la concordance et conjointement à jouer pleinement des complémentarités avec d'autres techniques voisines. Le "bon KWAC" repose fondamentalement sur les effets visuels d'alignements, parallélismes et correspondances, affinés et démultipliés avec le concept de zones que nous proposons de mettre en œuvre dans les concordanciers. Et la complémentarité affirmée articule les diverses formes de relevés d'attestation (KWOC, KWAC et KWIC redéfinis, et prolongés par le KWUT), comme elle motive l'intérêt pour des concordances d'auteur aux côtés de concordanciers bien conçus et utilisés avec méthode.

BIBLIOGRAPHIE

BÉHAR, H. 1997. La méthode d'Hubert de Phalèse, *Lexicometrica*, 0, 8 p.

BOURION, É. 2001. *L'aide à l'interprétation des textes électroniques*, Thèse de doctorat, Sciences du langage, Université de Nancy II.

CHOUEKA, Y. 1984. Conversationnel ou concordances imprimées : le problème de l'exploitation d'un gros corpus, *Informatique et Sciences Humaines*, 61-62, pp. 93-105.

LEBART, L., SALEM, A. 1994. *Statistique textuelle*, Paris, Dunod.

PINCEMIN, B., ISSAC, F., CHANOVE, M., MATHIEU-COLAS, M. 2006. Concordanciers : thème et variations, in J.-M. VIPREY (éd.), *Proc. of JADT* 2006, pp. 773-784.

RASTIER, F. 2001. *Arts et sciences du texte*, Paris, Presses Universitaires de France.

SALTON, G. & MCGILL, M. J. 1983. *Introduction to Modern Information Retrieval*, McGraw-Hill.

SÉKHRAOUI, M. 1995. *Concordances : Histoire, méthodes et pratique*, Thèse de Doctorat, Université de la Sorbonne nouvelle Paris 3 et Ecole normale supérieure de Fontenay Saint-Cloud.

L'ANALYSE FORMELLE DES EGODOCUMENTS DANS UN SYSTÈME INFORMATIQUE DE PRODUCTION DE RESSOURCES ÉLECTRONIQUES

Dominique TAURISSON
CNRS / EHESS

SOMMAIRE

Résumé : *Les Egodocuments sont une mine d'informations pour les historiens et les littéraires, mais leur richesse et leur surabondance constituent souvent un obstacle à leur bonne exploitation. Certains systèmes d'analyse assistée permettent déjà d'en faire des explorations partielles ; nous présenterons, ici, une expérience d'analyse formelle assez différente, mais complémentaire. Dans le cadre de travaux de production de ressources scientifiques électroniques utilisant un système informatique original et adapté (Arcane), les chercheurs rassemblent, organisent et analysent sources et documents de travail dans une base de données relationnelles considérée comme un « monde » : ils définissent les objets de ce monde (Sujets Personnes, Sujetslieux, SujetsInstitutions, SujetsOuvrages, etc), et formalisent les concepts et notions constitutifs des problématiques qui les intéressent sous forme de « relateurs » : par exemple, la rencontre, l'échange, le partage, le contrat, le déplacement, le transfert, etc. Les relateurs sont des expressions formelles composées d'un nom et d'une suite d'arguments : objets du monde, termes du métalangage, dates, modalités, etc. ; ils servent à établir des liens dynamiques entre les objets du monde et accroissent la combinatoire sémantique de ce monde. Les chercheurs parcourent leurs documents textuels et images, et peuvent décomposer jusqu'à la plus petite unité de sens les informations « intéressantes de leur point de vue » que ces documents enferment. Ils les enregistrent et les stockent dans les relations ad hoc, par instanciation de relateurs. Plusieurs relations peuvent décrire et être ancrées sur une même séquence de caractères, ou un segment d'image. L'exposé montrera comment la mise en œuvre de ce mode électronique d'analyse, génère une qualité et une grande profondeur de lecture : textes et images se dévoilent au chercheur dans toute leur richesse, leur complexité intricative, et leur feuilletage naturel (ce qui se voit et ce qui ne se voit pas), d'où une démultiplication du dialogue entretenu entre les documents et le chercheur. Des utilisateurs de disciplines différentes peuvent d'ailleurs explorer au moyen d'autres relateurs, correspondant à leurs propres domaines d'enquête, les mêmes matériaux, et produire ainsi une multiplicité d'interprétations et de points de vue. Les informations ainsi collectées et structurées peuvent être comparées, croisées, on peut en faire des analyses statistiques ou qualitatives et, également, en produire des représentations sous diverses formes graphiques. Ces informations, et les documents dont on les a extraites, sont exportables au format XML, à destination d'autres systèmes d'information. Outre l'intérêt immédiat de faciliter les travaux personnels des chercheurs et de les initier à de nouveaux modes de production de la connaissance, ce système me semble associé de deux façons à la problématique du colloque : - en permettant de développer un langage commun formalisé (objets, concepts) utilisable pour mettre en relation plusieurs « mondes » électroniques, et assurer communication et échanges entre eux, il repose, à nouveaux frais, la question de la mutualisation des recherches - en mettant en œuvre de nouvelles formes d'élaboration et de diffusion des savoirs, il s'inscrit naturellement dans le débat sur les accès à ces savoirs.*

1. Les egodocuments

De nombreux chercheurs s'accordent aujourd'hui pour considérer que les egodocuments[1], en tant que gisements exceptionnels d'informations, représentent une des voies d'accès possible à l'étude des relations interpersonnelles et des espaces relationnels des acteurs sociaux[2], autrement dit, de la sociabilité comprise comme « l'ensemble des formes concrètes, des modalités, des structures et des processus de mise en communication et de socialisation des individus dans une société donnée »[3].

Or, pour l'Age moderne, les correspondances, catégorie par excellence des egodocuments, sont encore assez régulièrement exclues des sources dites « du for privé »[4]. Elles ne sont pas toujours reconnues comme des documents sérieux et *vrais* du privé et donc fiables et exploitables pour la recherche, ni comme des sources neutres dès lors qu'elles déformeraient le « réel » puisqu'« elles obéissent à des règles de savoir-vivre et de mise en scène de soi par soi ».

Par ailleurs, traditionnellement absorbés par l'étude et la publication de corpus des personnages remarquables de cette époque, les chercheurs jugent souvent suspectes les correspondances produites par des individus modestes ou plus obscurs : des lacis de multiples arrangements, aux structures encore peu lisibles, où les agents sont tour à tour sollicités, recommandés, protecteurs, ceci au sein d'un réseau compliqué de relations de natures diverses aussi bien familiales, qu'administratives, commerciales ou politiques, souvent sans intérêt littéraire, parfois singulièrement dépourvues de pensée, à la fois considérables et lacunaires, truffées de mentions émiettées relatives à la vie quotidienne (« monnaie courante » de la sociabilité), où l'on peine finalement à repérer des informations significatives.

Thomas Grosser soulève de son côté d'autres problèmes associés en général aux sources des historiens, et aux egodocuments en particulier : « De façon générale, on peut remarquer que la « banque de données » de l'historien est beaucoup plus *limitée et fragmentaire* que celle du sociologue ou du socio-psychologue. Car ces derniers peuvent analyser des phénomènes actuels à l'aide d'expériences et « créer » eux-mêmes leurs données. Au contraire, l'historien *ne peut pas* trouver pour chaque variable et chaque cas les données empiriques correspondantes à des relations de facteurs complexes, car il est dépendant d'informations qui lui sont transmises de façon fragmentaire et pas toujours dépourvues de *déformations*. Cependant, on peut *reconstruire une mosaïque* plus ou moins vraisemblable, qui permet d'éclairer les grandes lignes du processus et que les théories modernes permettent de modéliser[5]. ». Outre l'intérêt de pointer précisément une question récurrente sur laquelle nous allons revenir, Grosser esquisse une des pistes de travail que nous avons scrupuleusement suivie : « la reconstruction de la mosaïque ».

2. Technologie de lecture

Pour réfléchir aux deux catégories fondamentales que T. Grosser dégage : données extraites et données créées, on peut se reporter à l'ouvrage de Maurizio Gribaudi, *Espaces, temporalités, stratifications. Exercices sur les réseaux sociaux*[6], où un processus classique de création de données par des historiens est justement décrit.

L'équipe de Gribaudi a demandé à des personnes choisies de faire le compte rendu, dans des cahiers d'enregistrement déjà structurés, de leurs contacts quotidiens directs et indirects, de leur parcours biographique et de leurs modèles de référence relationnels.

Les enquêtés détaillent, par exemple, leurs rencontres dans une période donnée : le moment de chaque rencontre, sa durée, son contenu, le lieu, le nombre et les personnes présentes, la description de la personne rencontrée, ainsi que les circonstances de leur première rencontre, etc. La collecte de données génère donc ici une production textuelle dont le contenu apparaît comparable à celui accumulé dans de nombreux egodocuments.

Dans ce cas (les cahiers d'enregistrement), la production informative peut être immédiatement formalisée : il y a structuration *a priori* de l'enquête, sous la forme d'une grille, d'un modèle, qui

---

[1] Néologisme proposé dans les années 1960 par le chercheur néerlandais Jacob Presser pour désigner les textes où l'on parle de soi.
[2] Beaurepaire et taurisson, 2003.
[3] Agulhon, 1986.
[4] Foisil, 1986.
[5] Grosser, 1992.
[6] Gribaudi, 1988.

sert tout à la fois, à la réalisation de l'enquête elle-même, à l'analyse et à l'interprétation des résultats.
Dans l'autre cas, celui des chercheurs en sciences humaines confrontés à des egodocuments, il apparaît nécessaire d'élaborer, *a posteriori* cette fois, une sorte de technologie de lecture pour repérer et extirper des informations hétéroclites de leur gangue textuelle et les structurer selon des modèles adaptés.
Dans les deux cas, l'objectif est d'analyser les informations collectées ou « dénichées », la place et le rôle des agents qu'elles concernent, qu'il s'agisse de personnes, de groupes, de lieux ou de productions, ainsi que les mécanismes de leurs interrelations.
Pour abonder dans le parallèle établi entre données créées et données extraites, l'équipe de Gribaudi, elle-même, montre que les données créées sont à prendre avec autant de prudence que les données des historiens. Les participants d'une enquête ont, en effet, des réactions diverses qui peuvent aller jusqu'à la parasiter, certains s'en servent comme d'une source d'inspiration, d'autres sont des « répondants à problème » qui accumulent retards, oublis, réticences, négligences et qu'il faut accompagner pendant des mois pour les aider à gérer la rédaction de leur cahier… Par ailleurs, rien n'assure que les données récoltées soient complètes, objectives, et *vraies*, même dans le cas des « répondants sans problèmes », le facteur psychologique n'étant pas vraiment quantifiable. M. Gribaudi écrit d'ailleurs qu'on obtient finalement : « trop de nuances, trop de renseignements, trop de vies individuelles… ».
Même si l'on convient unanimement de la nécessité d'élaborer une technologie de lecture adaptée capable d'aider les chercheurs à réaliser des micro-analyses de leurs sources, il ne faut pas perdre de vue que s'il est fort *probable* que les informations contenues dans des egodocuments peuvent leur apprendre quelque chose des comportements de ceux qui les ont rédigés, il est, en revanche, fort improbable de découvrir dans *un* corpus (la correspondance de X, par exemple) toutes les données significatives concernant le scripteur : notamment toutes les personnes qu'il connaît, tous les lieux auxquels il est lié. *A fortiori*, les liens suggérés par la liste de ses correspondants et par le contenu des lettres échangées ne constituent certainement pas l'ensemble des relations de cette personne, ce qui fait qu'il n'est guère possible d'expliquer les stratégies individuelles par les seuls témoignages des individus concernés.
L'abondance des documents et des renseignements qu'ils contiennent se conjugue donc paradoxalement à « un manque », à une absence, que l'instrumentation informatique mise au point devra compenser par la possibilité donnée aux chercheurs de croiser et de confronter différents corpus et sources documentaires.

3. Contexte problématique

La plupart des informations enfouies dans les egodocuments de l'Age moderne parvenus jusqu'à nous concernent des *SUJETS*, personnes, lieux, institutions, groupes (familles, organisations), productions écrites, artistiques, idées, et des *LIENS directs ou indirects* tissés entre ces Sujets.
Or, pour circonscrire l'espace relationnel d'un Sujet, notamment dans le cadre de travaux sur la sociabilité, il est nécessaire de savoir pour une personne, qui et qu'est ce qu'elle connaît, quels usages elle fait de ces connaissances, quels profits elle en retire, comment elle les a acquises, comment elle les entretient à travers les différents états de sa vie ; pour un lieu ou une institution, qui a eu des contacts directs ou indirects avec ce lieu (qui l'a visité ou habité), quand, pourquoi ; pour un groupe, qui en fait partie, depuis quand, à quelles opérations ses membres ont-ils été associés ; pour une production, qui l'a créée, qui l'a lue ou vue, dans quel endroit, qui en a parlé, etc.
Ce qui équivaut pour un Sujet à un ensemble variable de liens évolutifs inscrits dans différents contextes, et donc à un espace relationnel sans cesse reconfiguré.
En outre, chaque lien est à étudier en fonction de tous les autres liens, de la position relationnelle et des « attributs » (propriétés, connaissances) des Sujets, ces trois facteurs s'influençant les uns les autres, d'où évidemment une forte interaction entre identités, comportements et espaces relationnels (espace relationnel du Sujet et espaces relationnels des autres Sujets en relation avec lui) : dans le cas des relations interpersonnelles, interviennent donc les caractéristiques de chaque individu concerné, les contacts qu'ils entretiennent, ainsi que les règles ou codes de comportement en vigueur à un moment précis. Autrement dit, s'intéresser à leurs parcours signifie étudier en même temps les espaces relationnels dans lesquels ils s'inscrivent, tout en rendant compte de la structuration de ces espaces.

Ce qui revient à définir l'espace relationnel d'un individu à un instant « t » comme la concrétisation momentanée de son parcours à l'intérieur de l'espace social*,* son « capital relationnel » (humain et social)[1], et son espace relationnel global comme la somme de ses identités, de ses trajectoires et des liens enregistrés à tous les instants « t » de sa vie.
Si l'espace relationnel complet d'une personne paraît dès lors singulièrement inaccessible, il semble, en revanche, possible d'isoler et d'observer certains des segments qu'elle mobilise dans le cadre de stratégies personnelles à des moments précis de sa vie, et de comparer les circonstances et les modes de leur activation, leur évolution dans le temps, éventuellement d'étendre, au travers d'une démarche comparative, ces observations à d'autres corpus et à d'autres personnages.
Par extrapolation, on pourrait tenter de cerner l'intérêt d'un Sujet Lieu en circonscrivant son « espace relationnel » à différentes périodes, en notant les configurations de personnes et de pratiques (activités) associées à ce lieu : les rapports périodiques ou ponctuels entretenus par différents types de Sujets avec ce lieu, et leurs transformations, les personnes qui s'y sont rencontrées, qui en ont parlé, qui y sont revenu, les œuvres qui y ont été créées, celles qui ont pris ce lieu comme espace de narration, etc. Le travail consisterait à élaborer une typologie des lieux à partir de l'étude des fonctions relationnelles dont ils sont habituellement le théâtre, et dont on établirait une nomenclature.
Dans le même esprit, et pour revenir aux individus et aux groupes, il s'agirait de se concentrer sur leurs comportements et leurs agissements vus à travers quelques grands types de liens relationnels, assimilés à des formes élémentaires de la sociabilité : la recommandation, la rencontre, le partage, l'échange sous ses différentes modalités (don, prêt, achat-vente, etc.), méthode qui permettrait d'agréger et d'observer des Sujets d'une façon moins egocentrée que dans les études de réseaux traditionnelles.
Suivre ces différentes pistes rapidement esquissées, revient *à préconiser* pour l'analyse des egodocuments, notamment des correspondances, des cheminements scientifiques qualitatifs et compréhensifs, et pas seulement quantitatifs, de manière – à améliorer les modèles destinés à instancier et à expliquer, à partir de la multiplicité des pratiques et des micro-histoires, les formes et les invariants des comportements sociaux, et – à vérifier l'hypothèse que c'est de « l'ensemble des choix individuels que résultent certains processus macroscopiques ». Et bien entendu, nous souscrivons pleinement à ce qu'Alain Bensa rajoute, à savoir qu'il faut évidemment se doter *des outils et des moyens méthodologiques* pour étudier dans les textes les événements microhistoriques singuliers que nous venons de décrire, et pour les relier à d'éventuels systèmes plus englobants de données et de significations[2].

4. Instrumentation et analyse formelle

Nous cherchons donc à déceler ce qui se voit et ce qui ne se voit pas, ou mal, ce qui n'est pas explicite, à percevoir et à décomposer les feuilletages et les intrications d'information, puis à isoler et réorganiser les unités de connaissance ainsi récoltées pour multiplier les angles de vue et les échelles d'observation ; enfin, à offrir à différents types de lecteurs des accès pratiques pour mener leurs propres recherches : une façon de laisser parler les textes et de mettre en ordre la grande masse des éléments dispersés, combinés et hétéroclites qu'ils contiennent[3].
Pour toutes ces raisons, et dans l'objectif à moyen terme de favoriser les investigations croisées dans les egodocuments, on a conçu un instrument électronique (ArcaneWriter) et proposé un modèle de structuration de la connaissance (Arcane)[4]. L'un et l'autre associés, permettent aux chercheurs d'enregistrer et d'étudier, sous forme de bases de données propres, mais connectables, des documents multimédias (en particulier des egodocuments textuels) dans leur transcription philologique et dans leur matérialité initiale (images des manuscrits) : au fil d'une lecture approfondie, le chercheur enrichit ses documents au moyen d'une indexation systématique sous forme de notions-concepts très finement spécifiées ; dans le cas des correspondances de l'époque moderne et d'études sur la sociabilité : la circulation, la transmission, l'appartenance, l'échange, le transfert, ou l'aide,… (autant de signes de relations entre personnes, lieux, institutions, idées, productions, et objets) : il note tout ce qui peut caractériser et relier les individus,

[1] Lemercier, 2005.
[2] Bensa, 1996, p. 40.
[3] *Alter histoire*, 1991, p. 182.
[4] Lochard et Taurisson 2000 ; Lochard et Taurisson 2001.

les lieux, les groupes, les idées et les productions, ainsi que toute indication relative à la datation, à l'intensité et à la durée des relations enregistrées.
Nous verrons qu'en rendant ainsi possible la manipulation simultanée de différents contextes, échelles et points d'approche, un outil d'observation-enregistrement comme Arcane, sorte de microscope pour molécules textuelles, peut doublement aider le chercheur – à construire des visions formalisées et des répliques même fragmentées des modes de vie et des environnements des acteurs sociaux, – dans le cadre d'études sur les interrelations, à envisager la mise au jour d'une sorte de typologie des « formes de la sociabilité » considérées non plus comme des processus singuliers et isolés dans une époque et un espace donnés, mais comme des phénomènes stables que l'on découvre inexorablement à l'œuvre, selon des modalités à étudier, dans différentes formes de sociétés. L'hypothèse étant bien que ces catégories relationnelles sont génériques, peut-être invariantes, et qu'elles participent, il s'agit encore de comprendre dans quelle mesure, « à la construction et à la fortification des liens sociaux ». En 1986, Maurice Agulhon avait déjà proposé de choisir un cadre d'étude circonscrit et d'étudier l'ensemble *des formes de sociabilité,* qu'elles soient institutionnelles ou privées, à un moment donné pour mieux en percevoir les continuités et les interférences[1], et il avait été l'un des premiers, dès les années 60[2], à travailler sur le comportement privé et quotidien, « susceptible d'analyses relationnelles », alors même qu'il ne disposait d'aucun outil informatique pour mener à bien cette tâche.

5. Analyse formelle et *Relateurs*

Le principe est donc simple : chaque chercheur ou équipe de recherche nourrit sa propre base de données structurée compatible avec l'architecture *Arcane*, tout en conservant sa logique scientifique, ses problématiques de recherche, et ses contraintes éditoriales. Il accepte les principes et modalités de rassemblement de l'information préconisés, et en conséquence, de mettre à disposition du public et de ses collègues chercheurs un certain nombre des données signées, qu'il a rassemblées et traitées.
Quant à la mise en œuvre de la méthode d'analyse formelle des textes exposée succinctement plus haut, et définie dans le paradigme *Arcane*, elle peut être décrite et caractérisée par quatre opérations principales faciles à adapter aux problématiques particulières de chaque chercheur :
- Le chercheur définit, en rapport avec ses objets de recherche, les notions-concepts qu'il se propose de repérer dans les documents composant son corpus. Il isole les unités sémantiques qui constituent ces notions-concepts de manière à pouvoir les formaliser et les implémenter dans un système d'information électronique sous forme de *relateurs*.
Les *relateurs* sont des expressions « prédicatives » formées par une suite de valeurs ordonnées qui permettent d'établir des liens entre les Sujets (Lettres, Personnes, Lieux, Produits, Idées, Dates, Evénements) et des expressions métalinguistiques, selon des combinaisons formelles.
- Le chercheur sélectionne et enrichit systématiquement des séquences de caractères ou d'image, en les liant à un ou plusieurs de ces *relateurs*.
- La quantité considérable d'informations élémentaires (unités de sens) ainsi collectées et enregistrées peuvent elles-mêmes, selon des mécanismes intellectuels et électroniques, être mises en relation, recombinées et interprétées selon des problématiques claires et propres à chaque utilisateur. Ces mécanismes simples permettent de développer de nombreux autres traitements, et d'enrichir la sémantique du « monde » commun au chercheur et à son lecteur en définissant des requêtes pour l'interroger. L'exécution dynamique de ces requêtes provoque la sélection et l'affichage des objets vérifiant ces requêtes. Avec ces résultats, il est possible d'éditer des listes et de réaliser de nombreuses représentations graphiques dynamiques *in situ* dans l'environnement Arcane, ou par export vers d'autres applications mieux adaptées aux demandes spécifiques des chercheurs.
Un système comme Arcane est aussi bien équipé des outils nécessaires pour représenter les connaissances, et les textes d'où elles ont été extraites, que pour les éditer simultanément sur supports électronique (*via* XML) et papier (*via* TeX), nous en reparlerons plus loin.

[1] Agulhon, 1986.
[2] Agulhon, 1963.

6. Instrumentation et production de ressources

6.1. Croisement des sources et mutualisation

Il est évident que la manipulation d'un tel système allait naturellement conduire ses utilisateurs à vouloir étendre leurs recherches à d'autres corpus et sources d'information, à entreprendre des échanges et des croisements, opérations difficiles à mener jusqu'à maintenant, autrement que manuellement ou par de lourdes recherches automatisées, compte tenu des outils électroniques disponibles et de la difficulté de mettre en œuvre des travaux collectifs concertés et interopérables. Cette perspective devenue aujourd'hui réaliste constitue de notre point de vue, nous l'avons déjà souligné, une condition incontournable à un travail sérieux mené à partir de sources egodocumentaires, en particulier à cause de leur manque intrinsèque de fiabilité.

Plusieurs chercheurs et équipes expérimentent déjà la méthode d'analyse textuelle décrite plus haut[1], d'où la nécessité de maintenir une cohérence sémantique dans le choix des concepts manipulés et dans leur formalisation. L'objectif étant d'élaborer des méthodes *générales* d'analyse formelle des contenus textuels adaptées, en particulier, à l'étude des relations interpersonnelles et plus généralement des processus inscrits dans les egodocuments, il s'agit d'établir collectivement un ensemble minimum commun de *relateurs* utilisables dans les différents projets éditoriaux et/ou de recherche des utilisateurs ; ces *relateurs* constituant eux-mêmes un outil de production de ressources partageables.

Dans l'idéal, il serait souhaitable que quels que soient ces projets, et en dépit de leur diversité, il soit possible aux différentes bases de données qui les portent de « communiquer » de façon rationnelle et dynamique : or, pour établir des bases de connaissances disponibles et accessibles à tous, capables de dialoguer entre elles, il faut s'entendre sur ce qu'on enregistre et sur la manière de l'enregistrer. C'est pourquoi un intense effort de normalisation est mené au sein du groupe Arcane pour définir un métalangage commun (une grammaire) pour structurer, baliser, nommer, etc. les données.

Cela veut dire, notamment, élaborer collectivement des listes de types de Sujets avec leurs descripteurs (personnes, lieux, institutions, thèmes, productions…), de *relateurs* adaptés à l'étude de problématiques voisines ou convergentes (voir plus haut), de genres de documents, d'identificateurs universels, et d'énumérations descriptives[2].

En ce qui concerne les recherches sur les formes de sociabilité, on entrevoit déjà bien les bénéfices importants qu'elles pourraient retirer de la mutualisation et du croisement des sources, ne serait-ce que, nous l'avons déjà dit, pour compléter et confronter les informations manipulées. Quant à l'étude de l'instanciation des concepts fondamentaux qui relèvent de la sociabilité, comme la rencontre ou l'aide, elle trouverait les plus grands avantages dans la mise à disposition des chercheurs de très nombreuses occurrences comparables collectées dans des corpus de différentes périodes, et dans plusieurs types de documentation : contextes des rencontres, contraintes, processus, déroulements, résultats, etc. ; contextes des aides reçues, demandées, négociées, remerciées, modes de sollicitation, conséquences, effets, évolution des acteurs, etc.

6.2. Analyse formelle et projet éditorial

Ces opérations de mise en réseau de travaux individuels et de normalisation des méthodes d'enquêtes dans les sources textuelles vont naturellement de pair avec la volonté d'associer étroitement l'analyse de ressources documentaires multimédias à leur diffusion dans la communauté scientifique (éditions sous multiformats) et à la production simultanée de résultats de recherche ; il ne suffisait donc pas de mettre à la disposition des chercheurs une instrumentation *ad hoc* fortement adossée à leurs pratiques de recherche, mais il fallait l'adapter aux mécanismes qui ponctuent traditionnellement le circuit de l'édition : autrement dit, il fallait inventer un double outil d'écriture et de publication.

Le chercheur devait pouvoir mettre en œuvre, au sein de sa propre base de données et de façon autonome, différentes entreprises allant de l'écriture, à la recherche pure, jusqu'à la publication multisupports de documents de base (textes, images fixes et animées, son) et de connaissances scientifiques (papier, Web, CD-ROM), opérations respectant évidemment toutes les règles éditoriales préconisées par la communauté scientifique (notes de commentaire, notes critiques, mise en page élaborée, index et bibliographie cumulatifs, etc.).

---

[1] Voir Taurisson, 2005, *Actes des Journées d'étude sur l'instrumentation Arcane.*

[2] Voir Taurisson, 2005, *Actes des Journées d'étude sur l'instrumentation Arcane.*

Tout cela a été rendu possible par la mise au point d'un paradigme qui articule le couple auteur-lecteur autour d'un *monde* de connaissances structuré que l'auteur élabore et que le lecteur parcourt pour le comprendre : un ensemble réticulé de connaissances scientifiquement validées et architecturées dans lequel, grâce à un métalangage commun, les auteurs et les lecteurs peuvent agir collectivement[1], l'édition au sens traditionnel ne constituant plus qu'une des phases d'un processus coopératif sans limites dans le temps et dans l'espace qu'on pourrait détailler ainsi :
- Définir l'architecture du monde à éditer comme instance d'une méta-architecture
- Construire le monde des Sujets d'intérêt du projet éditorial étroitement associé au programme de recherche
- Enrichir ce monde par des documents multimédias et des relations (occurrences des *relateurs*) pour le décrire, le structurer, l'illustrer, l'interpréter et le représenter
- Enfin, en extraire à des moments choisis des sous-ensembles cohérents pour produire et publier des livres électroniques[2].

6.3. Accessibilité des ressources

L'accessibilité des ressources est aujourd'hui un problème central pour la diffusion de la connaissance par internet et de nombreux travaux y sont consacrés.
Au contraire des heuristiques qui favorisent des traitements en aval, au moment de la publication, traitements lexicologiques et linguistiques notamment, des outils d'écriture comme Arcane cherchent à donner aux auteurs les moyens de penser, de concevoir et de réaliser des accès intentionnels aux ressources qu'ils produisent, de manière à les cartographier conceptuellement, à en repérer les points d'entrée, et à baliser à l'intention de leurs différents types de lecteurs des navigations logiques dans l'hypergraphe des données ainsi constitué. Les *relateurs* jouent un rôle particulièrement important dans l'élaboration de cette sémantique de communication.

BIBLIOGRAPHIE

AGULHON, M. 1963. Les Associations, confréries religieuses et loges maçonniques en Provence orientale à la fin de l'Ancien Régime, *Actes du Congrès national des Sociétés savantes* ; section d'histoire moderne et contemporaine, 87, pp. 73-86.
AGULHON, M. 1986. La sociabilité est-elle objet d'histoire, *Sociabilité et société bourgeoise en France, en Allemagne et en Suisse, 1750-1850,* sous la dir. d'E. François, Paris, Recherche sur les Civilisations, pp. 13-22 (*Travaux et Mémoires de la Mission historique française en Allemagne*).
Alter histoire : essais d'histoire expérimentale, 1991, Daniel S. Milo et Alain Boureau avec Hervé Le Bras, Paul-André Rosental [*et al*], Paris, Les Belles Lettres.
BEAUREPAIRE, P.-Y. et TAURISSON, D., (éds.) 2003. *Les ego-documents à l'heure électronique : nouvelles approches des espaces et des réseaux relationnels*, Actes du colloque tenu du 23 au 25 octobre 2002 (CNRS & Université de Montpellier), Montpellier, SerPub, 2003, 554 p. (édition en libre consultation sur le Web : <http://www.univ-montp3.fr/arcanews/egodoc/>).
BENSA, A. 1996. De la micro-histoire vers une anthropologie critique, *Jeux d'échelle. La micro-analyse à l'expérience*, J. Revel (dir.), Paris, Gallimard, Le Seuil, 1996, pp. 37-70.
FOISIL, M. 1986. L'écriture du for privé, in P. Ariès et G. Duby (sous la dir.) *Histoire de la vie privée* de, Paris, Seuil, tome III (*De la Renaissance aux Lumières*), pp. 331-369.
GRIBAUDI, M. 1998. *Espaces, temporalités, stratifications. Exercices sur les réseaux sociaux*, Paris, EHESS.
GROSSER, T. 1992. Les voyageurs allemands en France. Etudes de cas et perspectives d'analyse, in J. Mondot, J.-M. Valentin et J. Voss (sous la dir.), *Allemands en France, Français en Allemagne 1715-1789. Contacts institutionnels, groupes sociaux, lieux d'échanges*, pp. 209-235.
LEMERCIER, C. 2005. Analyse de réseaux et histoire, *Revue d'histoire moderne et contemporaine*, 52-2, avril-juin 2005, pp. 88-112.
LOCHARD, E.-O. et TAURISSON, D. 2000. The World according to Arcane : an operating instrumental paradigm for schoolarly edition, *Proceedings* of the international conference

[1] Lochard, 2005, « Les documents électroniques dans le système d'écriture et d'édition Arcane » <http://hal.ccsd.cnrs.fr/ccsd-00016064>.

[2] Voir sur le Web l'édition du *Journal de Corberon (1775-1780)* réalisée avec Arcane dans les conditions décrites : *Web passif* : http://egodoc.revues.org, et *Web actif* : http://www.univmontp3. fr/arcanews/egodoc/ ; voir aussi Taurisson, 2004.

organized by the Constantijn Huygens Instituut and the Free University Amsterdam, Den Haag, 7-8 décembre 2000, Berlin, Weidler Buchverlag, 2002, pp. 151-162.
LOCHARD, E.-O. et TAURISSON, D. 2001. Le monde selon Arcane, *Le Document au XXIe siècle*, *Cahiers Gutenberg*, 39-40, mai 2001, pp. 89-105.
LOCHARD, E.-O. et TAURISSON, D. 2002. Correspondances, réseaux, édition électronique, *La Plume et la Toile. Pouvoirs et réseaux de correspondance dans l'Europe des Lumières*, textes réunis par P.-Y. Beaurepaire, Arras, Artois Université Presses, pp. 171-192.
LOCHARD, E.-O. 2005. Les documents électroniques dans le système d'écriture et d'édition Arcane <http://hal.ccsd.cnrs.fr/ccsd-00016064>
TAURISSON, D. et BEAUREPAIRE P.-Y. (éds.) 2000. Edition électronique du *Journal (1775-1781)* de Marie Daniel Bourrée, chevalier de Corberon <http://www.univ-montp3.fr/arcanews/egodoc/>.
TAURISSON, D. 2004. *Le journal de Corberon* sur le Web : édition numérique ou édition électronique ?, Actes du colloque *La numérisation des textes et des images : techniques et réalisations*, Université de Lille 3, 16 et 17 janvier 2003, textes réunis et édités par Isabelle Westeel et Martine Aubry, CeGes-Lille 3, 2004., pp. 47-62.
TAURISSON, D. (éd.) 2005. *Actes des Journées d'étude sur l'instrumentation Arcane*, organisées les 17 et 18 novembre 2005 par J. Boutier et D. Taurisson, SHADYC (EHESS CNRS, UMR 8562). <http://www.egodoc.revues.org/journeesArcane/>

EN QUOI LES ANALYSES PSYCHOLINGUISTIQUES PEUVENT-ELLES CONTRIBUER À L'ÉLABORATION DE SYSTÈMES DE RECHERCHE ET DE REPRÉSENTATIONS DES CONNAISSANCES ?

Martine CORNUÉJOLS
Chercheur associé à l'équipe MoDyCo – Université Paris X

SOMMAIRE

Résumé : *L'étude concerne l'organisation des réseaux sémantiques verbaux et imagées. La différenciation de ceux-ci est mise en évidence par la constitution de corpus d'associations imagées et verbales et l'analyse des types de liens associatifs. Le principe organisateur en terme de relations associatives de type catégories sémantiques est remis en cause. Une structuration en terme de situation, caractéristique et catégorie sémantique est proposée.*
L'analyse plus fine selon l'appartenance aux entités naturelles ou artificielles est réalisée en rapport avec les connaissances neuropsychologiques connues en terme de déficits catégorie spécifique.
Des perspectives dans le domaine de la gestion des documents, de l'indexation multimodale et du web sémantique sont envisagées.

1. Introduction

L'étude de l'organisation des réseaux sémantiques en mémoire humaine et de l'activation des significations ou des concepts par les perceptions visuelles peut servir de base pour élaborer des systèmes de représentation sémantique pour le traitement automatique du langage, en particulier pour la constitution d'ontologies et de lexiques et également des systèmes de recherche d'information (web sémantique, …) en contribuant à l'élaboration d'indexation. En effet, il paraît tout à fait important d'obtenir une bonne adéquation entre la représentation mentale de l'utilisateur final qui fait la requête pour sa recherche d'information et la représentation d'informations élaborée par le concepteur du système de recherche d'information. Cependant, la majorité des études sont axées sur le lexique et le langage verbal. Or les systèmes d'information évoluent vers le multimodal. On peut donc s'interroger sur la transposition des principes de sémantique verbale dans la sémantique imagée.

La psycholinguistique apporte un nouveau savoir sur l'organisation sémantique et comporte un volet expérimental de recueil de données concernant les réseaux associatifs et une partie modélisation pour construire une représentation des connaissances qui sera utilisable pour l'élaboration de systèmes d'information.

Les recherches exposées dans cet article portent sur la constitution des corpus verbaux et imagés correspondant aux associations sémantiques présentes en mémoire [Cornuéjols, 1999], à leur analyse aboutissant à l'élaboration d'une typologie des liens associatifs et à la mise en évidence de principes organisateurs (en terme de situation, caractéristiques et catégorie sémantique), à la mise

en relation des items verbaux et imagés pour modéliser les réseaux sémantiques, à la détermination des thèmatisations émergentes et à la comparaison avec des corpus textuels de type journaux d'actualité [Ferret & Cornuéjols, 1998] et de type roman [Reza & Rastier, 1999].
Ces recherches visent à contribuer à mettre au point des systèmes de recherche d'information basés sur les mots sémantiquement proches aux mots clés (approche itérative de la requête), à la modélisation sémantique à la base du web sémantique et également à la constitution de systèmes d'indexation multimodaux qui ne soient pas nécessairement basés sur la dénomination ou la description des images, comme c'est le cas actuellement, mais également sur la proximité sémantique et l'interrelation des réseaux imagés et verbaux.

2. Normes associatives verbales et imagées

2.1. Constitution des normes

Pour déterminer si le réseau associatif imagé est différent du réseau associatif verbal, une des méthodes est de comparer le corpus d'associations verbales et d'associations imagées. La littérature ne faisant mention d'aucune norme d'associations imagées existante, nous avons entrepris d'établir une norme d'associations imagées pour 285 items. Pour pouvoir la comparer avec la norme d'associations verbales et sachant que les normes d'associations verbales existantes ne comportaient pas uniquement des items concrets imageables, nous avons établi la table d'associations verbales correspondant aux dénominations des entités représentées par les images de la norme associative imagée. Ces données sont obtenues par une tâche d'association libre. Le dépouillement des résultats se fait en listant les occurrences citées par chacun des 134 participants face à l'item inducteur (soit environ 40000 données par norme) et à compter le nombre de fois (dénombrement et classement) où chaque occurrence a été citée sur l'ensemble des participants.
L'analyse globale comparative montre que les normes verbale et imagée divergent pour 77,26% des associés sémantiques cités. Une analyse plus fine sur l'associé majoritairement cité uniquement, montre que les normes associatives diffèrent dans 60,21 % des cas. Le recouvrement de l'associé majoritaire sur les deux normes est donc d'environ 40%.
Pour donner une idée sur un exemple concret, l'image du « zèbre » évoque spontanément dans la tête des personnes l'idée de « savane », alors que le mot « zèbre » lorsqu'il est lu, évoque « rayures » qui est une des caractéristiques de l'entité.
Tous ces résultats confortent l'idée de réseaux associatifs sensiblement différents pour l'image et pour le mot. Ces réseaux bien que différenciés sont intimement liés.
Ces corpus ont été validés par des expériences d'amorçage sémantique entre mots et images se basant sur les normes associatives ainsi établies. Un effet d'amorçage similaire est obtenu entre images et mots associés, sur la base de la norme associative imagée précédemment établie, et entre entités verbales sur la base de la norme associative verbale. En d'autres termes, l'image peut évoquer automatiquement le mot associé, tel qu'établit dans la norme associative imagée, mais ne peut évoquer automatiquement le mot que l'on supposerait associé, si l'on se base sur la norme associative verbale. La norme associative verbale ne permet pas de rendre compte des liens sémantiques entre images et entités verbales.

2.2. Proposition d'une typologie des liens associatifs

La grande dispersion des résultats nous a incité à proposer une typologie des liens associatifs. Dans tous les modèles de mémoire sémantique [Dubois, 1972 ; Denhière, 1975 ; Ehrlich & Tulving, 1976] on retrouve l'idée de catégories, dimensions ou composantes sur lesquelles les éléments stockés peuvent être classés et entretiennent entre eux un certain système de relations. La force associative et la nature des relations « associant-associés » reflètent certaines propriétés de la mémoire sémantique.
Une typologie en neuf points a été proposée : situation (spatiale, temporelle, événement), catégorie sémantique (super ordonné, classe, instance), attribut ou caractéristique, fonction (ex. : « couteau » - « couper ») ou produit (ex. : « miel » pour « abeille »), objet naturellement associé (ex. : « selle » pour « cheval »), association de type linguistique (ex. : « chapeau »-« melon »), symbole (ex. : « cœur » - « amour »), nul ou sans réponse (révèle la difficulté d'association), illisible (l'impondérable des expériences papier-crayon).

2.3. Analyse des normes associatives par rapport à cette typologie

La classification des associations sémantiques des normes verbale et imagée a été opérée par trois juges de façon indépendante. Le degré d'accord a été évalué par un coefficient de corrélation de Bravais – Pearson à .78 ($p<.001$).

Une analyse à deux niveaux peut être réalisée :

- l'une sur l'associé majoritaire uniquement ;
- l'autre sur l'ensemble des associés cités face à un item.

Les différences entre les tables d'associations verbales et imagées sont significatives pour les catégories typologiques Situation, Instance, Attributs, Fonction, Nul, Illisible.
Les associés de type Instances, Attributs et Fonctions sont plus souvent donnés en réponse à des stimuli verbaux qu'à des stimuli imagés. À l'opposé, les stimuli présentés sous forme d'images engendrent plus de réponses de type Nul ou Illisible, ce qui laisse supposer que la tâche d'évocation d'une image mentale est plus ardue que l'association simple entre mots.
Ce résultat conforte l'idée que les structures associatives correspondant aux images et aux mots ne sont pas similaires.

2.3.1. Le rôle du contexte

Le résultat le plus important de cette étude est la constatation que, quel que soit le niveau d'analyse, la présentation d'une image évoque préférentiellement la situation (ou contexte spatial ou temporel ou encore événementiel), dans laquelle l'entité représentée par l'image est généralement rencontrée. Le contexte le plus souvent évoqué est un contexte spatial, plutôt que temporel. Le sujet évoque préférentiellement une localisation plutôt qu'un événement situé dans le temps. La présentation d'une entité lexicale de type substantif évoque elle préférentiellement une caractéristique qui spécifie l'entité représentée par le mot. Ce qui voudrait dire que quand on voit un objet isolé, spontanément on le re-situe dans un contexte, un environnement, alors que lorsqu'on lit sa dénomination, on évoque spontanément un trait qui le caractérise.
Les associations majoritaires sont donc de type situation (spatiale essentiellement et événementielle) et objets associés ou contigus. Ceci corrobore les résultats de Yeh & Barsalou (2000) qui montrent que les concepts apparaissent en situation et non abstraits de celles-ci. Ceci explique aussi pourquoi les concepts peuvent apparaître sous différentes formes selon les situations, où chaque forme contient les propriétés pertinentes à la spécificité de la situation (Yeh & Barsalou, 2000). Donc il semblerait que les principes organisationnels de la mémoire sémantique diffèrent en fonction du contexte. Par exemple, *un piano* est un instrument de musique, mais dans un contexte situationnel de déménagement, son poids devient un trait saillant. Le fait que chaque item présenté isolément évoque le même associé cité par environ 60 % des personnes suggère que le stockage du poids ou de la saillance des traits reste le même à l'intérieur de la population, mais que le contexte peut faire varier ce facteur. Cela suggère donc que la saillance des propriétés des items isolés peut être différente de celle des items en situation. Le fait que le contexte apparaisse spontanément dans les associés majoritaires peut inciter à penser que le contexte lui-même fait partie des traits qui définissent un objet. Le rôle du contextuel dans l'organisation des associés sémantiques est analysé également en fonction du type d'item source.

2.3.2. L'organisation en catégories sémantiques

Un autre résultat important issu de cette analyse est la constatation que les liens associatifs de type catégorie sémantique et instance, qui révèlent des liens de hiérarchie taxinomique, ne représentent que 10,23% pour les associations imagées et 21,84 % pour les associations verbales. Ce résultat suggère donc que l'organisation taxinomique n'est pas prédominante dans le réseau associatif verbal ou imagé, alors même que la littérature le pose comme principe organisateur prépondérant dans l'organisation conceptuelle [Collins & Quillian, 1969 ; Rosch, 1975 ; Sartori & Job, 1988].
De plus, la majorité des liens taxinomiques sont de type instance ou classe équivalente, ce qui suppose donc que le sens privilégié est du général vers le spécifique et non de l'instance (pourtant représentée par l'item source) vers la catégorie plus globale.

2.3.3. Organisation globale tripartite

L'étude des associations montre que l'on peut regrouper celles-ci en deux grands groupes : les associations qui spécifient l'objet par opposition aux éléments qui le replacent dans un contexte.

Dans le premier groupe sont regroupées les classes instance, attributs et fonction, alors que dans le deuxième se retrouvent les classes situation, items associés ou contigus, et de catégorie (super ordonné, comme contexte hiérarchique).
En ce qui concerne les *associations verbales*, 30,58 % des associations produites spécifient les entités considérées, alors que 66,74 % les contextualisent.
Pour ce qui est des *associations imagées*, 11,15 % des associations spécifient l'entité tandis que 88,85 % des associés fournis les contextualisent. La contextualisation, bien que majoritaire dans les deux tables d'associations, est fortement accentuée dans le cas des images, puisqu'elle apparaît dans presque 90 % des cas.
Pour ce qui est des associations verbales, dans plus d'un quart des cas l'associé produit vise à spécifier l'entité dénommée. Ces résultats confirment que les caractéristiques permettant de différencier les entités sont plus associées aux mots qu'aux images.
Si on dissocie les catégorisations sémantiques, qui sont un principe organisateur à part entière, on peut considérer trois modes d'organisation, le mode dominant étant la situation ou contexte (contexte essentiellement spatial ou événementiel), le deuxième étant la caractéristique (attribut ou trait, fonction, produit, …) et le troisième, la catégorie sémantique.

2.3.4. Analyse en fonction du type de catégorie (naturel / artificiel) des entités présentées

Une analyse des associations par catégorie sémantique a été réalisée pour déterminer si la nature de la catégorie sémantique de l'amorce (ou item source) avait une incidence sur la cohésion du noyau associatif central. En effet le corpus d'items source comprend 114 entités naturelles (animaux, végétaux, insectes, …) et 170 entités artificielles (véhicules, outils, monuments, vêtements, instruments de musique, …). Les données neuropsychologiques proviennent de cas de patients présentant des déficits catégorie spécifique ou des déficiences sélectives dans certains domaines sémantiques (Hillis & Caramazza, 1991; Warrington & Shallice, 1984 ; Saffran & Schwartz, 1992). La configuration la plus fréquemment rencontrée est l'atteinte touchant les connaissances du domaine des entités vivantes, alors que la sémantique des catégories sémantiques artificielles semble préservée. Parfois la configuration inverse est observée (Hillis & Caramazza, 1991 ; Moss & Tyler, 1997 ; 2000 ; Sacchett & Humphreys, 1992 ; Warrington & McCarthy, 1983, 1987).
Un de nos objectifs, par l'analyse des normes associatives et des types de liens sémantiques, est de mettre en évidence les différences entre entités naturelles et artificielles et de rendre compte de ce qui peut expliquer les déficits catégorie spécifique. Notre question était de savoir quel type de lien associatif peut être altéré qui expliquerait les déficits sélectifs observés et d'ainsi mettre en évidence le type de relation qui pourrait être dominant pour chaque catégorie d'item.
Une première remarque est que cette étude ne révèle aucun effet catégorie sémantique dépendant. Quelle que soit la catégorie de l'item source, en moyenne, environ 63 % des sujets ont cité le même pool d'items associés.

	"SITUATION"				"CARACTERISTIQUE"					"CATEGORIE"			
	sit	Ass-o	evt	act	Attr	prod	fct	sym	cont	Cl eq	inst	Cl cat	CL
Entités naturelles													
moy %	**21**	**3**	**4**	**3**	**21**	**8**	**3**	**3**	**2**	**9**	**10**	**7**	**6**
ET	18	8	6	9	20	16	8	7	7	15	17	10	10
%	31				37					32			
Entités artificielles													
moy %	**11**	**11**	**6**	**5**	**9**	**7**	**10**	**1**	**8**	**7**	**16**	**3**	**6**
ET	18	19	9	11	15	17	16	6	20	14	26	7	14
%	33				35					32			

Sit: Situation (localisation) ; **Act**: Acteurs; **Evt**: Evénement ; **Ass Obj**: Objet Associé ; **Sym**: Symbole ; **Attr**: Attributs ; **Prod**: Produit ; **Fct**: Fonction ; **Cont**: Contenu ; **Cl equ**: Classe équivalente ; **Cl Inst**: Instance; **Cl Cat**: Nom de la Catégorie.

Tableau 1 : Pourcentage des 3 liens associatifs des 3 modes d'organisation entre entités naturelles et entités artificielles

Les résultats du tableau 1 montrent que les entités naturelles sont également reparties entre situation, caractéristique et catégorie.
Une analyse plus fine montre que les différences apparaissent entre catégories naturelles et artificielles en terme de situation (majoritaire pour les entités naturelles), et d'objets associés (prédominant pour les entités artificielles) pour ce qui concerne le mode de structuration de type situation ; en terme d'attribut (majoritaire pour les entités naturelles) et de fonction (logiquement plus associée aux entités artificielles) pour ce qui concerne les liens caractérisant ; en terme d'instance (majoritaire pour les entités artificielles) et les noms des catégories sémantiques (ou super ordonné) (majoritaires pour les entités naturelles) pour le mode de structuration de type catégorie.

Une analyse encore plus fine consiste à différencier les différents types d'entités naturelles et d'entités artificielles et de déterminer le type de lien prédominant selon le type spécifique d'items.

	SITUATION					CARACTERISTIQUE					CATEGORIE			
	Sit	Act	Evt	Ass-obj	sym	Char	prod	fct	sp	contents	Cl equ	Cl inst	Cl cat	Cl
WA	22.22	0	5.55	0	0	**33.33**	0	0	11.11	0	5.55	16.66	0	5.55
DA	8.33	0	0	0	0	16.66	**58.33**	0	8.33	0	8.33	0	0	0
B	**45.45**	0	0	0	0	27.27	0	0	0	0	0	9.09	18.18	0
I	**33.33**	0	0	0	0	11.11	22.22	0	0	0	22.22	11.11	0	0
SA	**80**	0	0	0	0	20	0	0	0	0	0	0	0	0
BP	17.64	0	0	23.52	5.88	**29.41**	0	5.88	0	0	5.88	0	5.88	5.88
UP	16.66	8.33	8.33	0	0	**25**	8.33	0	0	8.33	16.66	8.33	0	0
VF	**28**	0	0	4	0	20	12	0	0	0	12	12	4	0

WA: Wild Animals (animaux sauvages) ; **DA**: Domestic Animals (animaux domestiques) ; **B**: Birds (oiseaux) , **I**: Insects (insectes) ; **SA**: Sea Animal (animaux marins) ; **BP**: Body Parts (éléments du corps) ; **UP**: Universe Parts (éléments de l'univers) ; **VF**: Vegetables and Fruits (végétaux et fruits).
Sit: Situation; **Act**: Actors; **Evt**: Event; **Ass Obj**: Associated Object; **Sym**: Symbol; **Char**: Characteristic; **Prod**: Product; **Fct**: Function; **Sp**: Cultural specificity; **Cont**: Contents; **Cl equ**: Equivalent class; **Cl Inst**: Instance; **Cl Cat**: Name of Category.

Tableau 2 : Répartition des associés dans les catégories pour chaque type d'entité naturelle.

	SITUATION					CARACTERISTIQUE					CATEGORIE			
	Sit	Act	Evt	Ass-obj	sym	Char	prod	fct	sp	contents	Cl equ	Cl inst	Cl cat	Cl
Cloth	10.52	5.26	0	**36.84**	0	5.26	0	0	0	10.5	10.52	15.78	0	5.26
Vehi	**40**	0	10	30	0	0	0	0	0	0	10	0	0	10
Furn	5.88	0	5.88	**23.52**	0	0	17.64	11.76	0	5.88	17.64	11.76	0	0
Build	**25**	8.33	16.66	8.33	8.33	8.33	0	0	0	8.33	0	16.66	0	0
Gam	12.5	12.5	12.5	0	0	0	0	12.5	0	0	12.5	**37.5**	0	0
Mus	10	0	10	10	0	0	**60**	0	0	0	0	10	0	0
Cont	11.76	0	5.88	0	0	0	0	0	0	**52.9**	0	23.25	0	5.88
O S	0	0	12.12	9.09	0	3.03	15.15	15.15	0	6.06	6.06	**33.33**	0	0
Weap	0	0	0	**40**	0	0	20	0	0	0	20	20	0	0
Tools	0	13.33	0	**40**	0	0	0	0	0	0	6.66	13.33	6.66	20
other	8.69	4.34	4.34	13.04	4.34	0	4.34	13.04	8.69	0	8.69	**17.39**	0	13,04

Cloth: Clothes (vêtements) ; **Vehi**: Vehicles (véhicules) ; **Furn**: Furnitures; **Build**: Buildings (monuments) ; **Gam**: Games (jeux) ; **Mus**: Musical Instruments (instruments de musique) ; **Cont**: Containers (récipients); **O S**: Office Supplies (outils de bureau) ; **Weap**: Weapons (armes) .

Tableau 3 : Répartition des associés pour chaque type d'entité artificielle

Comme le montrent les tableaux ci-dessus, l'associé majoritaire dépend de la catégorie de l'objet. Ceci est à mettre en rapport avec les données de déficits neuropsychologiques.

2.3.5. Les thèmes évoqués par les associations sémantiques

Cette étude fait émerger une organisation des connaissances en termes de situation et de thématique. Plusieurs associés peuvent être regroupés par thème évoqué.

Par exemple, dans la norme associative verbale, pour « abeille » on retrouve le thème de la production de miel, le thème de la piqûre, …

1	ABEILLE	**miel**	48	34,78%	produit	s	fct
2	ABEILLE	ruche	20	14,49%	sit (lieu)	s	lieu
	ABEILLE	nid	1	0,72%	sit (lieu)	s	lieu
3	ABEILLE		15	10,87%	rien		
5	ABEILLE	guêpe	9	6,52%	cl (equ)	m	cat
	ABEILLE	frelon	1	0,72%	cl (equ)	m	cat
	ABEILLE	Maya	2	1,45%	cl (inst)	m	cat
	ABEILLE	insecte	3	2,17%	cl (cat)	t	cat
	ABEILLE	rayures	2	1,45%	car	a	struct
	ABEILLE	jaune	1	0,72%	car	a	struct
	ABEILLE	ailes	3	2,17%	car	a	struct
	ABEILLE	tache	1	0,72%	car	a	struct
	ABEILLE	bzzzzzzzz	2	1,45%	produit	a	struct
	ABEILLE	dard	1	0,72%	car	a	struct
4	ABEILLE	piqûre(s)	12	8,69%	evt	s	fct
	ABEILLE	butiner	2	1,45%	fct	s	fct
	ABEILLE	fleur	2	1,45%	obj-ass	s	lieu
	ABEILLE	prés	1	0,72%	sit (lieu)	s	lieu
	ABEILLE	chat	1	0,72%			?

Pour le même exemple dans la norme imagée :

ABEILLE	2	**miel**	29	21,17%	23,36%	prod
ABEILLE		pot de miel	3	2,19%		prod
ABEILLE	1	**piqûre**	31	22,63%	28,47%	Sit (evt)
ABEILLE		piqûre d'insecte	1	0,73%		
ABEILLE		piquer	2	1,46%		
ABEILLE		danger piqûre	1	0,73%		
ABEILLE		dard	1	0,73%		
ABEILLE		allergie au venin de guêpe	1	0,73%		
ABEILLE		oedème	1	0,73%		
ABEILLE		méchant	1	0,73%		
ABEILLE	3	ruche	13	9,49%	19,71%	Sit (lieu)
ABEILLE		nid	1	0,73%		
ABEILLE		essaim	2	1,46%		
ABEILLE		toile d'araignée	1	0,73%		
ABEILLE		grenier	1	0,73%		
ABEILLE		cave	1	0,73%		
ABEILLE		boîte transparente	1	0,73%		
ABEILLE		camping l'été	1	0,73%		
ABEILLE		confiture	1	0,73%		
ABEILLE		été	5	3,65%		Sit (t)
ABEILLE		prairie	1	0,73%		Sit (lieu)
ABEILLE	4	fleur(s)	10	7,30%		Sit (lieu)
ABEILLE		pollen	1	0,73%		
ABEILLE		bouton	3	2,19%		
ABEILLE		arbre	1	0,73%		
ABEILLE		insecte	5	3,65%	11,68%	Cl (cat)
ABEILLE		Maya	1	0,73%		Cl (inst)

ABEILLE	mouche	4	2,92%	
ABEILLE	ABEILLE	3	2,19%	
ABEILLE	guêpe	1	0,73%	Cl (equ)
ABEILLE	araignée	1	0,73%	
ABEILLE	moustique	1	0,73%	
ABEILLE	odeur	1	0,73%	
ABEILLE		1	0,73%	
ABEILLE	bzzzzzzzz	1	0,73%	
ABEILLE	homme	1	0,73%	
ABEILLE	horreur	1	0,73%	

2.3.6. Le cas particulier des associés linguistiques

Pour ce qui est des associations de type linguistique, elles sont très minoritaires et ne représentent que 3,67 % des associations. Cette classe, bien que spécifique du verbal, pose des problèmes de classification. Les associations de type linguistique pourraient être assimilées à la catégorie « objets associés ». Doit-on la classer comme une composante de la catégorie instance (ex. « Chapeau – melon » est une instance de « chapeau ») ou comme un objet (linguistique) associé (ex. « rouge - gorge » qui n'est pas une instance de « rouge » et qui ne définit l'objet que par l'association des deux items) ?

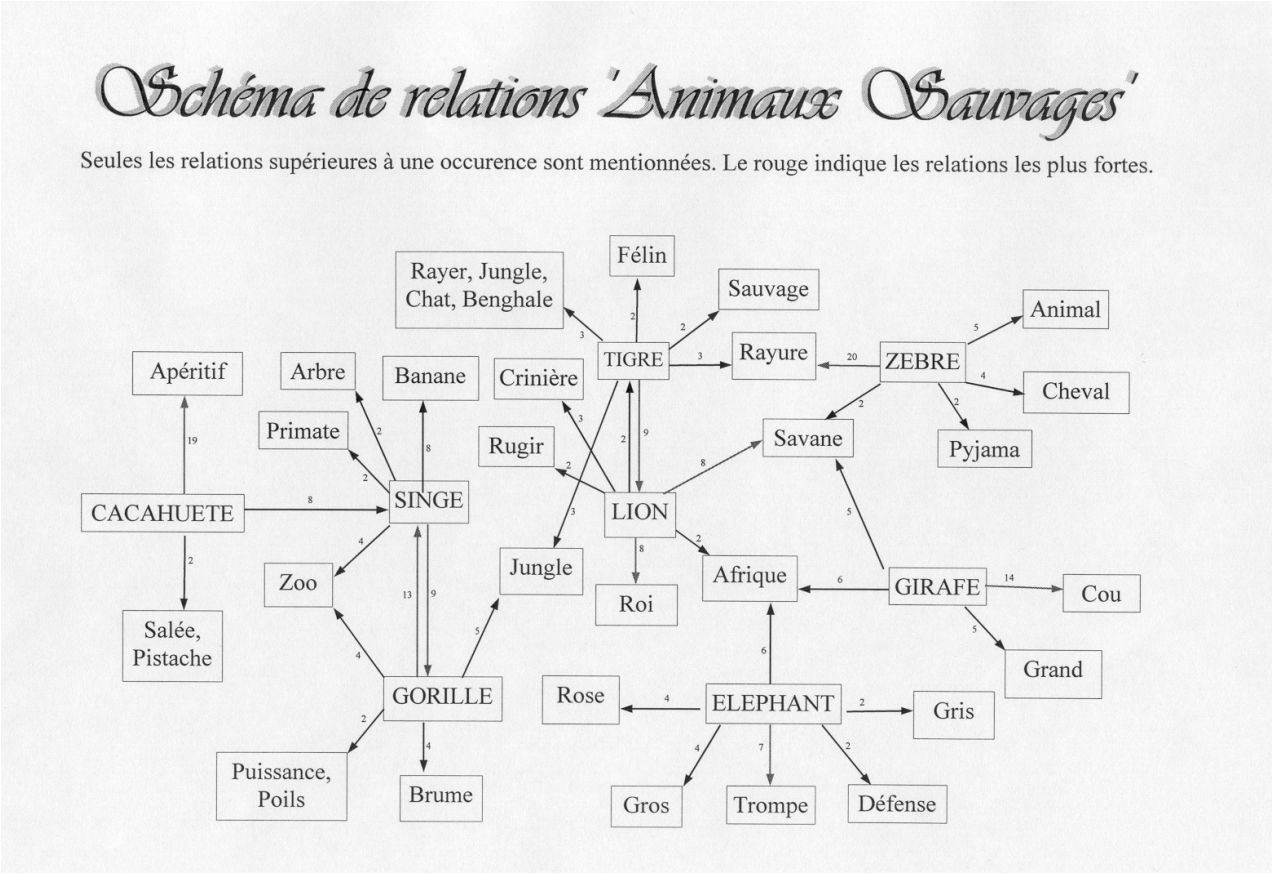
Les associations linguistiques montrent bien que ce n'est pas l'associé qui est catégorisé, mais bien l'association qui l'est.

Les données montrent que 30,58 % des associations linguistiques produites spécifient ou caractérisent l'entité (instancier), alors que dans 66,74 % des cas, elles la contextualisent (contexte linguistique).

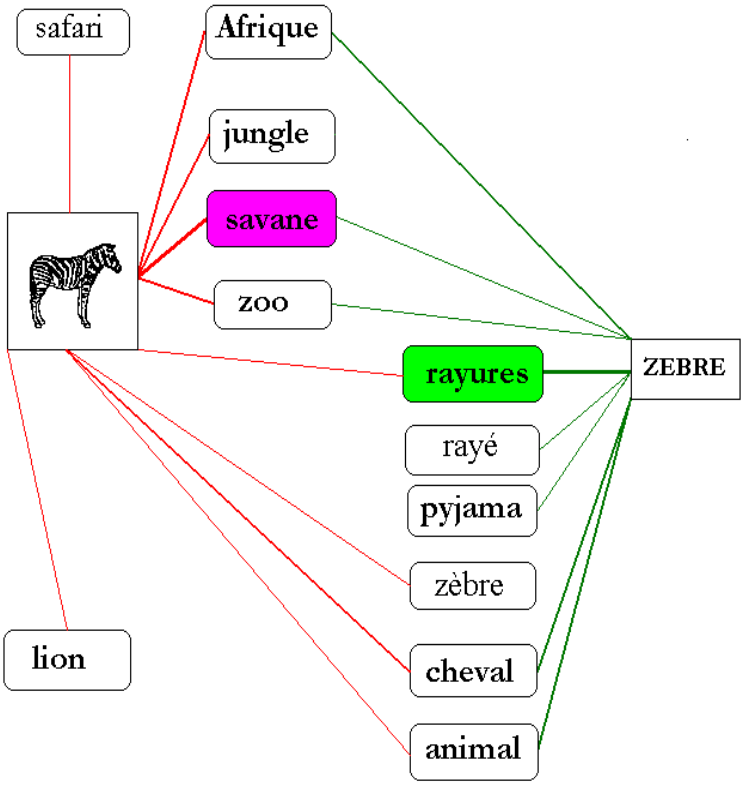
3. Des normes associatives aux réseaux sémantiques

La démarche que nous avons suivie a permis à partir de chaque item inducteur de percevoir le champ sémantique couvert par les associés cités. Une autre démarche possible est celle consistant à partir d'une notion commune de rechercher comment une notion se réalise dans différents signifiants.

La mise en relation des éléments lexicaux est donc opérée au niveau des catégories d'objets pour tenter de représenter les relations entre items appartenant à la même catégorie sémantique et de constituer ainsi les réseaux sémantiques des entités à la manière de Collins et Loftus (1975).



Nous avons également mis en relation réseau sémantique verbal et imagé.



4. Comparaison des associations sémantiques avec des cooccurrences de corpus textuels

Le postulat de base est que les relations associatives reflètent l'usage verbal. Les données d'associations de mots recueillis dans la norme associative verbale que nous avons constituée sont comparées aux associations (cooccurrences de mots) trouvées dans des contextes phrastiques issus de corpus textuels de type Le Monde [Ferret & Cornuéjols, 1998] et de type roman [Reza & Rastier, 1999]. Le pourcentage de recouvrement des associations verbales avec les cooccurrences textuelles est faible dans le cas du corpus issu du Monde car ce journal est très spécifique (politique, économique), alors qu'il est de 40% avec un corpus de type roman. Ceci accrédite la validité des normes associatives.

Les normes associatives imagées et verbales constituent des éléments de base utilisables dans les systèmes d'indexation d'images et de mots du web sémantique.

5. Perspectives pour la représentation des connaissances dans les systèmes artificiels, la recherche documentaire, le web sémantique, l'indexation multimodale, …

Une des problématiques communes que nous avons avec le domaine du traitement du langage naturel est celle de la nature et de l'organisation des connaissances et le formalisme adéquat pour les représenter.

Les associations reflètent l'organisation en mémoire sémantique mais également l'usage verbal qui en est fait en production écrite. Ces corpus peuvent être un bon indicateur des associations spontanées entre termes que vont faire les opérateurs utilisateurs d'un système de recherche d'information. Ils permettent d'évaluer les phénomènes de fréquence d'occurrence et de co-occurrence de mots qui révèlent certains aspects de la structure sémantique sous-jacente à l'usage verbal.

L'une des perspectives est utiliser ces réseaux associatifs conjointement à des formalismes de représentation sémantique des textes pour élaborer des graphes conceptuels qui permettent une indexation des documents plus efficace, et donc une recherche d'information plus efficace. Cela permet aussi de constituer des dictionnaires électroniques à partir des associations de mots permettant de mieux cerner la requête de l'utilisateur si le mot ou le groupe de mots le plus efficace pour sa recherche ne lui vient pas à l'esprit lors de la consultation du système. Par ces associations de mots et ces dictionnaires, le système pourra fournir des résultats plus proches de la demande réelle de l'utilisateur. Ceci est encore du domaine de la recherche et des perspectives d'avenir. Constituer ainsi des réseaux de mots-clés, plutôt que des requêtes sur de simples mots isolés pourrait être intéressant.

Le problème particulier des images en indexation est de savoir quel est le type de trait de l'image qui peut jouer le rôle de descripteur. La possibilité de faire porter l'indexation, non plus seulement sur des termes isolés, mais sur des nœuds dans un réseau de relations sémantiques, pourrait aider à obtenir des systèmes de traitement de l'information plus adaptés, non seulement pour cerner plus précisément le sens de la requête, mais également pour être plus en adéquation avec les représentations sémantiques des utilisateurs finaux.

BIBLIOGRAPHIE (TRÈS SUCCINCTE)

CORNUÉJOLS, M. 2001. *Sens du mot, sens de l'image*, Paris, Editions L'Harmattan.

Les références citées ci-dessus y sont répertoriées.

CORPUS ORAUX : LES *BONNES PRATIQUES* D'UNE COMMUNAUTE SCIENTIFIQUE

Olivier BAUDE
CORAL – Université d'Orléans EA 3850 / Délégation Générale à la Langue Française et aux Langues de France

SOMMAIRE

0. Introduction

Les problèmes juridiques liés à la diffusion des corpus oraux ont été l'occasion d'une démarche originale adoptée par une communauté scientifique ouverte à un travail pluridisciplinaire. Cette démarche a comporté plusieurs étapes. Une lecture croisée des textes juridiques par les linguistes et les juristes a permis de repérer les problèmes. Les chercheurs ont ensuite accepté d'expliciter leurs pratiques au regard de la législation. Cette étape fondée sur la réflexivité a permis d'élaborer des propositions pour de bonnes pratiques partagées par la communauté scientifique et de repérer des aspects juridiques qui posent des difficultés dans l'état actuel du droit.
Ce travail s'est concrétisé par la rédaction de l'ouvrage *Corpus oraux, guide des bonnes pratiques 2006*[1]. Rédigé par un groupe de travail constitué de linguistes, juristes, informaticiens et conservateurs, cet ouvrage a pour vocation explicite, d'éclairer la démarche des chercheurs, de repérer les problèmes et les solutions juridiques et de favoriser l'émergence de pratiques communes pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux.
Le résultat de ce travail interdisciplinaire ouvre les portes d'une réflexion sur les pratiques des chercheurs en sciences sociales et leurs relations aux données, à l'heure de l'exploitation et de la diffusion en masse de celles-ci.

1. Contextes pour une diffusion de la recherche

1.1. La linguistique de corpus et l'oral

Depuis plus de 30 ans le domaine de la linguistique de corpus s'est considérablement développé autour des corpus écrits, aussi bien en ce qui concerne la masse des données disponibles que l'élaboration d'outils de traitement automatique de celles-ci. La situation est totalement différente pour les corpus oraux[2]. Pourtant, les toutes nouvelles technologies en matière de stockage, de diffusion mais aussi d'exploitation des enregistrements sonores, couplées aux outils (transcriptions synchronisées sur le signal, annotations, etc.) ouvrent des perspectives prometteuses pour les études sur les corpus de langues parlées. De nombreux corpus ont été constitués ou sont en

[1] *Corpus oraux, guide des bonnes pratiques 2006*, Paris, CNRS éditions.
[2] Pour plus de commodités et selon l'usage, nous utiliserons les termes *corpus oraux* comme termes génériques définissant des collections ordonnées d'enregistrements de productions linguistiques orales et multimodales.

cours de constitution et leur diffusion pose des problèmes juridiques et éthiques que la communauté scientifique doit prendre en charge. Pourquoi et comment ?

1.2. Une politique de diffusion

Depuis 1982 et la loi pour la recherche et le développement technologique en France[1], la diffusion des résultats fait partie des missions des chercheurs. Plus récemment, la déclaration de Berlin signée par la plupart des Directeurs Généraux des Établissements Publics à caractère Scientifique et Technologique (EPST) le 22 octobre 2003 plaide pour la constitution de bases de connaissances en libre accès[2]. Enfin, les programmes de numérisation patrimoniale comprennent un volet de valorisation des ressources numérisées (cf. texte de Lund de 2001 prônant la mise en place des standards d'interopérabilité).

1.3. Les initiatives de mutualisation

Cette dernière notion de standards d'interopérabilité se retrouve dans différentes initiatives internationales (TEI, groupe de travail ISO TC37 SC4 pour la gestion des ressources linguistiques, protocole d'échange OAI, norme ANSI/NISO Z39.50, projet Open Language Archive Community, etc.) ainsi que dans des choix techniques (utilisation du langage de balisage XML par exemple). Dans le même cadre de valorisation de la recherche et de mutualisation des ressources, le CNRS s'est doté, en 2005, d'une direction de l'information scientifique, et développait un an plus tard des centres de ressources numériques.
Dans le même temps des laboratoires de recherche lançaient différentes initiatives pour la diffusion et l'accessibilité des corpus oraux (Base Clapi du laboratoire Icar[3], projet Corpus Oraux de l'EPML 50[4], programme Archivage du Lacito[5], constitution de grands corpus disponibles comme le projet Phonologie du Français contemporain[6], C-oral-Rom[7], etc.).

1.4. Le Guide des bonnes pratiques

C'est dans ce contexte que la Délégation générale à la langue française (direction du ministère de la culture) et le CNRS ont constitué un groupe de travail pluridisciplinaire qui a pour mission de favoriser la collecte et l'exploitation de corpus oraux.
Ce groupe de travail comporte des linguistes experts et des chercheurs de "terrain" porteurs de projets actuels, des représentants des fédérations de laboratoire du CNRS, des juristes, des représentants des grands organismes de conservation sous la tutelle du Ministère de la Culture et des juristes de ces institutions. L'objectif premier était de permettre un travail en commun sur un objet scientifique, de favoriser sa conservation et surtout sa diffusion (diffusion auprès de différentes équipes de recherche mais aussi auprès d'un public plus large). Or, il est très vite apparu que les aspects juridiques étaient les premiers obstacles à la diffusion de l'oral transcrit (qui est propriétaire de quoi ? Qui est responsable de la diffusion ? Quelles sont les autorisations à recueillir ? Qu'en est-il du droit d'auteur ?, etc.). Enfin, ce travail sur les aspects juridiques a très vite été lié à une réflexion sur l'éthique du chercheur et l'occasion d'une démarche réflexive sur ses méthodes.
Dans un premier temps, le groupe de travail s'est orienté vers l'élaboration par la communauté scientifique "de bonnes pratiques" avec les contraintes suivantes : premièrement il n'existe pas de réponses juridiques simples à l'exploitation de l'oral et à la transcription des données et deuxièmement les solutions passent systématiquement par un travail réflexif sur la démarche du chercheur, seul moyen pour qualifier le statut des enregistrements et les objets exploités. Les "bonnes pratiques " consistent donc à clarifier les questions juridiques, mais aussi - et c'est là un point fondamental - à porter une réflexion sur le travail scientifique des linguistes dans le respect d'une éthique validée par la communauté scientifique.

---

[1] Art 5 de la Loi n°82-610 du 15 juillet 1982 modifiée d'orientation et de programmation pour la recherche et le développement technologique de la France, aujourd'hui art. L 111-1 du code de la recherche. *JO* du 16-07-1982, p. 2273 et ss.
[2] Corpus oraux, Guide des bonnes pratiques op. cité, p. 36.
[3] Clapi-Icar http://clapi.univ-lyon2.fr
[4] EPML50 (ex Asila)
[5] Archivage du Lacito : http://lacito.vjf.cnrs.fr/archivage/index_fr.html
[6] PFC http://www.projet-pfc.net
[7] C-Oral-Rom 2005.

2. Aspects juridiques

D'une façon très schématique la réponse aux questions juridiques consiste à définir le statut juridique de l'objet "corpus" par ses conditions d'élaboration et sa composition, afin de procéder à la gestion contractuelle des droits des personnes concernées et de définir les responsabilités de ceux qui vont intervenir dans la vie du corpus (créateurs, hébergeurs, diffuseurs,…).

2.1. Définition de l'objet

Pour des raisons épistémologiques et techniques, la forme des corpus oraux est relativement complexe. Dans la majorité des cas les corpus oraux sont constitués :

- d'enregistrements (analogiques ou numériques) qui en cas de supports analogiques ont une durée de vie très courte avec une perte de qualité lors des migrations,
- de données contextuelles sur les locuteurs et la situation d'enquête qui peuvent être en partie des données personnelles (nom propre, profession, adresse, lieu, …),
- de transcriptions (sous la forme de fichiers indépendants ou permettant une synchronisation sur le signal ; transcription phonétique, orthographique, multilinéaire, etc.),
- d'annotations "secondaires" (informations sur les conditions de production des énoncés, précisions sur les phénomènes sonores tels que les rires et les bruits),
- d'annotations enrichies (étiquetage morphologique, syntaxique, annotations prosodiques pragmatiques, …),
- d'une documentation.

2.2. Domaines juridiques concernés

Pour définir le statut juridique de l'objet scientifique "corpus oral" et les droits des personnes concernées, il faut tout d'abord connaitre les conditions d'élaboration du corpus et de ses différentes composantes. Il s'agit ensuite de définir si le corpus est constitué d'informations du domaine public et/ou s'il est le produit d'une ou plusieurs créations intellectuelles susceptibles d'être protégées par le droit d'auteur. Il convient enfin de vérifier si le corpus contient des données personnelles qu'il faudra alors traiter. Ces statuts juridiques déterminés et les droits qui en découlent connus, il convient de s'enquérir des modalités de la gestion contractuelle de ces droits et de savoir si les titulaires de ceux-ci se sont prononcés sur les conditions de mise à disposition et de réutilisation des corpus - en apportant par exemple, leur consentement d'une manière formelle.

2.3. Diffusion scientifique et droit d'auteur

Seule une explicitation rigoureuse de la démarche du chercheur permet de savoir si un corpus est protégé par le droit d'auteur. Si tel est le cas, quels sont ces droits ?

Il convient de distinguer les droits patrimoniaux des prérogatives du droit moral. Les droits patrimoniaux se résument en un droit exclusif au profit de l'auteur (ou des titulaires) ou des ayants droit (bénéficiaires d'une cession, héritiers...) d'autoriser ou d'interdire la reproduction ou la communication au public de l'œuvre protégée. Quant aux prérogatives du droit moral, toujours attachées à la personne physique créatrice de l'œuvre protégée, elles sont au nombre de quatre : le droit de divulgation, le droit de repentir et de retrait, le droit à la paternité et le droit au respect de l'œuvre. En réalité, il existe une possibilité intermédiaire où les corpus protégés par le droit d'auteur peuvent être mis en libre accès dans le cadre d'une licence accordée par les titulaires de droits autorisant l'utilisation et l'exploitation des résultats (c'est le cas des Creatives commons). Sans être dans le domaine public, ces corpus sont – de par la volonté de leurs créateurs – libres d'accès et d'utilisation. Néanmoins, si les créateurs peuvent renoncer à exercer leurs droits patrimoniaux, il ne leur est pas possible de renoncer à leur droit moral qui reste imprescriptible.

3. Éléments pour de *bonnes pratiques*

3.1. Expliciter la démarche du chercheur

Les objectifs scientifiques, liés à la constitution, à l'exploitation, à la conservation et à la diffusion des corpus oraux sont très diversifiés, et le respect de ceux-ci, ainsi que leur hétérogénéité, impliquent que soit reconnue la diversité des démarches qui peuvent être adoptées par les chercheurs et par les utilisateurs ultérieurs de ces corpus.

Le Guide des bonnes pratiques n'a pas vocation à contraindre cette démarche en prescrivant une méthodologie type, mais souhaite fournir toutes les informations nécessaires au repérage des points juridiques et éthiques « sensibles ». Seule l'identification précise et détaillée des éléments

de la situation en jeu et notamment de la forme des données et de leurs supports, des pratiques de terrain, mais aussi des différentes étapes du traitement, permet d'apporter à la fois des éléments de réponses juridiques correspondant à la situation, et une évaluation des « risques » éventuels. Enfin, une analyse réflexive sur la démarche liée à la constitution et aux traitements des corpus oraux est le premier élément de l'élaboration d'une éthique reconnue par l'ensemble d'une communauté scientifique.

3.2. Le recueil de consentement

Le geste éthique le plus classique de la démarche du chercheur-enquêteur est le recueil de consentement du témoin. En réalité cette pratique est peu maîtrisée et souvent réduite à un formulaire de demande d'autorisation qui évoque en une phrase "le cadre d'un programme de recherche". Or sans informations préalables précises la demande d'autorisation n'a pas d'objet ni de sens. Pour que cette autorisation soit pertinente il conviendrait de concevoir le recueil d'un consentement "éclairé" qui démontre que le signataire est informé des finalités de la recherche et des conséquences à son égard d'une participation au projet.

Dans le cadre du recueil de données et notamment d'enregistrement pour des corpus oraux, le consentement devrait tenir compte de l'adéquation au destinataire (les informations fournies, pour être comprises doivent être adaptées aux compétences de compréhension du destinataire), et de l'explicitation des finalités de l'enquête (qui toutefois ne doivent pas renforcer le paradoxe de l'observateur en pointant l'objet de l'observation).

De plus, les explications sur le projet scientifique, doivent être complétées par des informations précises comme par exemple : les responsables de l'enquête et leur affiliation institutionnelle, ainsi que les financeurs ; une adresse de contact, les personnes qui auront accès aux données et qui travailleront sur elles, la façon dont les données seront anonymisées, le fait que les données seront transcrites selon des conventions particulières, la façon dont les données seront archivées une fois l'enquête terminée, les modalités d'accès aux informations relatives au projet et concernant tout particulièrement les données/analyses faisant référence à la personne (possibilité d'accès aux fichiers et informations concernant tout particulièrement la personne), les droits de la personne, notamment le droit de rétractation, les risques éventuels ainsi que les retombées positives, morales ou matérielles, de l'étude.

Enfin, le consentement devra préciser l'objet de la demande : les actions effectuées par les chercheurs dans le cadre du projet, les formats et les conditions de l'enregistrement, les conditions de diffusion des données et des résultats, les contextes de diffusion des données et des résultats.

Il est à noter que les formes de l'autorisation ne sont pas imposées par le législateur et qu'une demande orale enregistrée peut être valide et même parfois indispensable.

Sur le plan juridique, la collecte de données sensibles sans recueil de consentement est possible à la condition particulière que les données soient anonymisées dans un très bref délais. La procédure d'anonymisation est également très importante pour obtenir l'accord des témoins *a fortio*ri dans le cas d'une diffusion des données primaires.

3.3. L'anonymisation

Les pratiques actuelles des chercheurs en terme d'anonymisation se réduisent la plupart du temps à une opération de masquage d'un nom propre, d'une adresse ou d'un numéro de téléphone. Afin de vérifier la validité de ces pratiques et d'en définir les modalités, il convient de reposer avec précision la question légale qui est celle de l'impossibilité d'identifier des personnes. En effet, l'objectif est de protéger la vie privée des personnes enregistrées en dépersonnalisant les données, ce qui a amené le législateur à ne pas réduire cette identification à la simple présence de données nominatives.

Ainsi, si techniquement l'anonymisation consiste au remplacement ou au codage des données sensibles par des éléments neutres selon les supports concernés (remplacement par un blanc ou un pseudo à l'écrit, par un bip dans les fichiers sons et par floutage des visages sur les enregistrements vidéos), il serait erroné de penser que cette solution ne demande pas une expertise plus approfondie des risques d'exploitation d'éléments "dénommant".

3.4. Structure du corpus

Il existe d'autres possibilités que l'anonymisation par cryptage. Celles-ci reposent sur des limitations techniques prévues par la structure du corpus. La loi québécoise « concernant le cadre

juridique des technologies de l'information » propose de protéger l'anonymat non pas en modifiant les données, mais en limitant les possibilités de recherche, voire en les adaptant à la personne qui consulte la base selon des critères bien précis (sa profession, une autorisation, sa présence dans le fichier, etc.)
Cette dernière perspective offre pour la constitution et l'exploitation de corpus oraux la possibilité de faire coïncider les obligations légales avec les nécessités du travail de recherche. Toute donnée étant potentiellement sensible, une anonymisation systématique s'avère de plus en plus complexe ; elle peut même mettre en danger l'intérêt de certaines recherches. En effet, des détails concernant les personnes comme par exemple le nom, ou le lieu d'habitation peuvent constituer un élément important du corpus, ainsi que des résultats que l'on peut en tirer. C'est pourquoi la possibilité de ménager des niveaux d'accès selon des critères stricts (ex : chercheur ou non, présence d'autorisation, but de la consultation, etc.) semble une alternative efficace. Il existe d'autres procédés à inventer. En effet, l'article 11-2 de la nouvelle loi ouvre la possibilité de faire certifier des techniques nouvelles par la CNIL.

4. Conclusion

La démarche originale présentée ici a plusieurs intérêts. Outre le fait qu'elle offre les garanties d'une diffusion des corpus pour la recherche et pour d'autres finalités, elle impose une posture éthique aux collecteurs, utilisateurs et diffuseurs de corpus. C'est aussi l'occasion de porter un regard réflexif sur des pratiques et sur une démarche scientifique peu souvent explicitée. Enfin, il s'agit de permettre la constitution de corpus dont la mutualisation est la première étape d'une démarche scientifique rigoureuse qui ouvre les portes de l'analyse et de l'interprétation.

BIBLIOGRAPHIE

BAUDE, O. 2006. *Corpus oraux, Guides des bonnes pratiques, 2006*, CNRS-Editions et Presses Universitaires d'Orléans.

BAUDE, O. 2004. Les corpus oraux entre science et patrimoine. L'expérience de l'observatoire des pratiques linguistiques, in *Actes du Colloque international du GRESEC « La publicisation de la science »* (Grenoble), pp. 7-11.

BIBER, D. 1985. *Variations across spoken and written language*, Cambridge, CUP.

BIBER, D. 1999. *Longman Grammar of Spoken and Written English*, Londres, Longman.

BILGER, M. (dir.) 2000. Linguistique sur corpus, études et réflexions, *Cahiers de l'université de Perpignan*, Perpignan, Presses universitaires.

BILGER, M. (éd.) 2000. *Corpus, Méthodologie et applications linguistiques*, Paris, Champion.

BLANCHE-BENVENISTE, Cl. & JEANJEAN, C. 1987. *Le français parlé : transcription et édition*, Paris, Didier-Erudition.

CALLU, A. & LEMOINE, H. 2004. *Patrimoine sonore et audiovisuel français : entre archive et témoignage : guide de recherche en sciences sociales*, 7 vol., 1 CD-Rom, 1 DVD-Rom, Paris, Belin.

CAMERON, D., FRAZER, E., HARVEY, P., RAMPTON, M. & RICHARDSON, K. 1991. *Researching Language : Issues of Power and Method*, London, Routledge.

CONDAMINES, A. (éd.) 2006. *Sémantique et corpus*, Paris, Hermès.

CRESTI, E. & MONEGLIA, M. (éds.) 2005. *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam/Philadelphie, Benjamins.

CRIBIER, F. & FELLER, E. 2003. *Projet de conservation des données qualitatives des sciences sociales recueillies en France auprès de la « société civile »* rapport présenté à Madame la Ministre déléguée à la Recherche et aux nouvelles technologies, dactylogr. 2 vol. et http://www.iresco.fr/labos/lasmas/rapport/Rapdonneesqualita.pdf

ENCREVE, P., & FORNEL de, M. 1983. Le sens en pratique, *ARSS* 46, L'usage de la parole.

HABERT, B., NAZARENKO, A. & SALEM, A. 1997. *Les linguistiques de corpus*, Paris, A. Colin.

JACOBSON, M. 2004. Corpus oraux en linguistique de terrain, *Traitement Automatique des Langues*, 45/2, pp. 63-88.

JACOBSON, M. 2004. Les archives sonores au LACITO, *Bulletin de liaison de l'AFAS* 26 (http://afas.mmsh.univ-aix.fr/bulletin/Bulletin AFAS 26.pdf).

JOUTARD, P. 1979. Historiens, à vos micros. Le document oral, une nouvelle source pour l'histoire, *L'Histoire* 12, pp. 106-113.

KENNEDY, G. 1998. *An introduction to Corpus Linguistics*, Londres, Longman.

LABOV, W. 1972. *Sociolinguistic Patterns*, Philadelphie, University of Pennsylvania Press.

LEECH, G. 1992. The state of the art in corpus linguistics, Aijmer & Altenberg (éds.), pp. 8-29.

MONDADA, L. 1998. Technologies et interactions sur le terrain du linguiste. Le travail du chercheur sur le terrain. Questionner les pratiques, les méthodes, les techniques de l'enquête, Actes du Colloque de Lausanne 13-14.12.1998, *Cahiers de l'ILSL* 10 : 39-68.

MONDADA, L. 2006. Video recording as the reflexive preservation-configuration of phenomenal features for analysis, Knoblauch, H., Raab, J., H.-G. Soeffner, Schnettler, B. (éds.).

MONDADA, L. (à paraitre) La demande d'autorisation comme moment structurant pour l'enregistrement et l'analyse des pratiques bilingues, *Tranel*, Université de Neuchâtel.

QUÉRÉ, L. *et al.* (éds.) 1984. *Arguments ethnométhodologiques*, Paris, Centre d'Étude des Mouvements Sociaux, EHESS.

Recherches sur le Français Parlé 5 1984. Pourquoi le français parlé est-il si peu étudié ?

Revue Française de Linguistique Appliquée 1996. 1-2, 1999. IV-1.

SACKS, H. 1984. Notes on methodology, J. M. Atkinson & J. Heritage (éds.), pp. 21-27.

SHAFFIR, W.B. & STEBBINS, R. A. (éds.) 1991. *Experiencing Fieldwork : An inside View of Qualitative Research*, Londres, Sage.

SILVERMAN, D. (éd.) 1997. *Qualitative Research. Theory Method and Practice*, Londres, Sage.

SINCLAIR, J. 1991. *Corpus, Concordance, Collocation*, Londres, OUP.

SINCLAIR, J. 1996. *Preliminary recommendations on corpus Typology*, Technical Report, Eagles.

SINCLAIR, J. & COULTHARD, R. M. 1975. *Towards an Analysis of Discourse*, Londres, OUP.

« Speech Annotation and Corpus Tools », A special issue of *Speech Communication* 33, 1-2 2001. Steven Bird and Jonathan Harrington.

WELLAND, T. & PUGSLEY, L. (éds.) 2002. *Ethical Dilemmas in Qualitative Research*, Aldershot, Ashgate.

POUR UNE HERMÉNEUTIQUE NUMÉRIQUE : MÉDIATISER L'ACTIVITÉ D'ANNOTATION

Gaëlle LORTAL[a], Amalia TODIRASCU[b], Myriam LEWKOWICZ[a]
a CNRS Tech-CICO / Université de Technologie de Troyes b LILPA / Université Marc Bloch, Strasbourg

SOMMAIRE

Résumé : *Alors que les échanges médiatisés s'accroissent, le document numérique devient central, en particulier dans des activités distribuées. En effet, pour se comprendre, les participants d'un projet doivent partager un référentiel commun. Ce dernier se co-construit par des échanges dans lesquels les lecteurs transmettent leur interprétation sur des documents. Dans le cas des documents numériques, l'annotation marginale, c'est-à-dire la production de fragments de documents liés à un document et aidant à expliquer ce document, est encore peu soutenue. Le principe de l'annotation est par ailleurs largement utilisé dans le domaine du Web Sémantique, qui la définit comme une métadonnée. L'annotation est alors comparable à un index montrant le chemin vers l'information et permettant son accès. Selon nous, annoter n'est pas uniquement laisser une note esseulée dans la marge, mais c'est construire un réseau de sens autour du document, une interprétation.*

Nos travaux visent à intégrer ces perspectives en soutenant l'annotation comme moyen, à la fois de soutenir la négociation, de produire des fragments textuels, et aussi d'indexer finement des documents et fragments existants. Il s'agit donc de soutenir la discussion autour de textes élaborés ou en cours d'élaboration par un collecticiel qui permette la constitution de documents.

L'annotation médiatisée, définie ainsi, émerge dans diverses disciplines. En CMO[1], M. Marcoccia définit en effet un forum comme un « document dynamique produit collectivement et interactivement et dont la cohérence du contenu et de la forme est le résultat d'une gestion collective et coopérative ». Les posts du forum sont des fragments de documents reliés entre eux par leur indexation à un même thème, telles des post-its collés à un même document papier. Par ailleurs, A. Bénel souligne, lui, qu'un enjeu des bibliothèques virtuelles est d'offrir non seulement des sources documentaires, mais aussi leur apparat critique. C'est-à-dire un ensemble de lectures liées à la source qui négocient les sens du document et d'où peut naître une étude critique. Cet apparat est lui-même un document qui peut être critiqué et interprété à son tour.

Conserver ces traces de négociation permet de conserver la « logique de communication » qui mène à une création de document. Dans un environnement collaboratif, ces traces doivent être partagées et leur (ré-)interprétation peut appeler à l'esprit d'autres traces, créant des liens entre les documents et les fragments d'interprétation. Ces traces de lecture sont à la fois des traces de négociation et de conception de document. Pour l'étude de ces négociations, nous avons constitué un corpus en conception mécanique médiatisée fondé sur les échanges entre les différents concepteurs du projet. Le seul outil dans ce projet pour la communication asynchrone médiatisée est le mél. Nous avons choisi de conserver 27 méls (2200 « mots ») dont le corps est lié à un texte ou une pièce jointe. Le mél est en cela une annotation d'une pièce jointe, ou une annotation d'une autre annotation (mél en « réponse à », en « faire suivre »).

L'étude de ce corpus nous permet de tester nos hypothèses sur les fonctionnalités nécessaires à notre collecticiel. Cet outil permettra de médiatiser l'annotation, c'est-à-dire de mettre en place des techniques d'ancrage de fragments textuels, mais aussi des techniques d'organisation de ces fragments (indexation) pour permettre une réutilisation de ceux-ci dans une production de nouveaux

[1] Communication Médiatisée par Ordinateur.

documents. On constate alors que les annotations sont un élément de production de documents, permettant de nouvelles négociations et étant en cela des artefacts de négociation de frontières.

1. Introduction

Alors que le travail médiatisé se développe, les documents numériques s'installent dans les pratiques collaboratives. La médiatisation des pratiques d'édition et de publication permet plus de souplesse dans la gestion collaborative de documents. Le groupe qui élabore et partage des documents peut aujourd'hui les gérer en autonomie. Les problématiques de constitution et de classification de documents sont donc modifiées et les outils informatiques disponibles ne sont pas toujours adéquats. Pour comprendre ces pratiques collaboratives médiatisées et les assister au mieux, nous avons observé un projet collaboratif en conception mécanique dans le cadre d'une association aéronautique universitaire. À partir de ces observations, nous avons proposé un collecticiel qui propose de soutenir la gestion collaborative des documents.

Le projet observé a été mis en place entre le bureau d'une association aéronautique et une équipe de conception mécanique (chercheurs et techniciens). L'équipe d'ingénierie mécanique travaille de manière asynchrone et distribuée (ils sont localisés sur trois sites). Leur objectif est d'adapter un moteur automobile diesel pour un avion Delvion essence. La collaboration des différents participants est principalement portée par le partage de documents textuels et de plans. Par l'observation de leurs différents échanges pour l'élaboration de leur projet, nous avons constaté que l'annotation des documents est cruciale à la fois pour communiquer sur les documents et le projet et pour les re-contextualiser. Tout comme dans le champ de l'herméneutique, annoter n'est pas uniquement laisser une note esseulée dans la marge ; c'est construire un réseau de sens autour du document, une interprétation. Cette interprétation est partagée par d'autres lecteurs qui annotent à leur tour et mettent en place un cercle herméneutique (Gadamer, 1996).

Pour l'étude des pratiques annotatives, nous avons donc constitué un corpus fondé sur des plans, des maquettes et des documents numériques échangés par pièce jointe de mél. Ces documents sont la base de l'évolution du projet puisqu'ils expliquent ou représentent le travail à effectuer en production. Ils sont modifiés au fur et à mesure des échanges ou rencontres et il est important que les intéressés soient tenus informés des mises à jour et modifications des documents. Les membres de l'équipe de conception annotent les documents pour les modifier, pour informer les autres membres de l'importance d'une modification ou expliquer une telle modification, accompagnent leurs documents numériques d'une note dans un corps de mél, expliquent par un mél le contenu d'un document joint,… En clair, le passage au travail médiatisé et au document numérique, supplée l'annotation par différentes techniques agissant sur la communication et la révision de document. Le seul outil dans ce projet pour la communication asynchrone médiatisée est le mél. Nous avons observé environ 450 méls (environ 28 000 « mots ») dont le corps est lié à un autre texte (« réponse à »/RE : ; en « faire suivre »/FWD ; « mise en copie »/CC :) ou à une pièce jointe. Le mél est en cela une annotation d'une pièce jointe, ou d'une autre annotation. Ce corpus, selon la perspective théorique qui lui est appliquée devient une base au développement d'une fonctionnalité logicielle pour soutenir les pratiques médiatisées des concepteurs en mécanique. En effet, selon les perspectives de différentes communautés de recherche, l'annotation est un medium qui permet de communiquer, d'élaborer des documents ou de construire des classifications de documents. Ces perspectives mettent à jour différentes dimensions de l'annotation et impliquent des fonctionnalités informatiques différentes.

Nous allons maintenant présenter chacun des trois angles théoriques considérés, et, pour chaque angle théorique, décrire une partie du corpus représentatif d'une pratique d'annotation spécifique. Nous exposons ensuite les fonctionnalités mises en place sur cette observation. Tout d'abord, l'annotation en tant que fragment de conversation est découverte dans notre corpus à la lumière des recherches en Communication Médiatisée par Ordinateur et implique des fonctionnalités informatiques permettant l'échange et la négociation. L'annotation en tant que fragment de document, elle, vient de l'étude des exercices médiévaux et de l'herméneutique et donne lieu à des fonctionnalités d'édition collaborative. La dernière dimension d'étude applicable à notre corpus est issue du champ du Travail Collaboratif Assisté par Ordinateur et au Web Sémantique et engendre des possibilités de création de classification et d'indexation de documents et fragments de documents.

2. Un fragment de conversation

L'annotation est avant tout un élément de communication. Elle permet l'échange entre divers auteurs d'interprétations, d'opinions, de critiques,... En CMO[1], dans une perspective de constitution d'archive, M. Marcoccia (Marcoccia, 2001) définit un forum comme un « document dynamique produit collectivement et interactivement et dont la cohérence du contenu et de la forme est le résultat d'une gestion collective et coopérative ». Les « posts » du forum sont des fragments de documents reliés entre eux par leur indexation à un même thème. Ces posts servent à la négociation entre différents utilisateurs, à donner leurs opinions sur un thème. En cela, les annotations que nous avons pu observer sont très proches. Dans une perspective plus orientée sur les processus d'échanges (Marcoccia, 2004) définit aussi le forum comme un polylogue puisque plusieurs interlocuteurs participent à un fil de discussion. On constate que dans un contexte de travail collaboratif, l'annotation déposée sur un document partage ces attributs de conversation polylogale où des fragments de conversation sont liés à un thème. Dans cette perspective de CMO, (Labbe et Marcoccia, 2005) ont aussi mené une recherche sur la genèse du mél. (Labbe et Marcoccia, 2005) constatent que le mél partage des propriétés du « billet » tel que définit par (Haroche-Bouzinac, 2000) comme appartenant au genre du « dialogue épistolaire bref ». Cette forme de dialogue écrit se caractérise principalement par sa brièveté, son style informel, ainsi que des caractéristiques informationnelle, séquentielle et relationnelle fortes. Le billet est « adressé » précisément à un destinataire et contient un message qui établit une relation directe avec celui-ci. Par l'utilisation du mél ou de la note, nous constatons que le message relie fortement un auteur, un destinataire et un document. Cette relation est établie implicitement par le contenu du message ou plus formellement par l'ancrage physique du message attaché à un document, collé à un bouquet ou accompagnant une pièce jointe dans un mél. L'annotation est comparable à ce type d'élément permettant de relier des documents soit dans leur ensemble (objet à part entière), soit par une partie de leur texte (citation d'une partie d'un document). Dans un cadre de travail collaboratif, l'annotation possède aussi ces caractéristiques : elle est adressée, liée à un ou des document(s), utilisée pour la planification d'une activité dans le groupe (un post-it collé à un dossier pour informer de la suite des actions à effectuer sur ce document), ou encore pour l'échange d'opinion à propos d'un document partagé (une note marginale argumentant sur une partie d'un document). L'annotation est un support à la communication entre plusieurs interlocuteurs et permet la construction d'une conversation dans un groupe. Suivant ces observations, nous considérons l'annotation comme un fragment conversationnel. Nous considérons que c'est sa dimension principale dans la mesure où toute annotation porte intrinsèquement un message et un lien. Un annotateur passe une information à un destinataire et crée un lien entre son opinion et un document. Il met aussi en relation plusieurs interlocuteurs par son commentaire. Annoter signifie alors communiquer via des fragments de texte.

2.1. Un corpus d'annotation conversationnelle

De cet éclairage théorique sur l'annotation pour la conversation, nous pouvons décrire des échanges tels que (1), (2), (3) et (4). Saturnin est chef de projet et fait le lien entre le Conseil d'Administration (CA) de l'association qui commande le moteur et les groupes de mécaniciens et bénévoles qui conçoivent, produisent et installent les pièces. Désiré est mécanicien - pilote et a en charge la pose du moteur et les tests de l'avion. Félicien est un ingénieur qui dirige une petite équipe de technicien dont fait partie Adelphe. Adelphe travaille plus particulièrement sur des commandes émanant de concepteurs tels que Léon.

Dans cette série d'échanges, Saturnin essaye d'obtenir les pièces nécessaires au montage du moteur. Félicien lui demande de suivre la procédure normale où le technicien (Adelphe) fabrique certaines pièces bien que lui-même ait les outils nécessaires (fraise acier) disponibles. On voit ici au fur et à mesure des réponses que Saturnin négocie le fraisage par Félicien mais que celui-ci ne répond que par la négative mettant en avant la distribution préconisée des rôles. Les thèmes et reprises de thèmes sont mis en gras.

[1] Communication Médiatisée par Ordinateur

(1)

From: Saturnin To: Félicien
Subject: Delvion : Bride de l'échappement et pièces du train avant pour le 13/02 dernier délais

Bonjour,

Désiré a pris contact avec moi, il aurait besoin de la bride de l'échappemment et des pièces du train avant pour le 13/02 au plus tard, **est ce envisageable ?.** Merci de me répondre au plus tôt.

Bonne journée.

Saturnin

(2)

From : Félicien To : Saturnin

Bonjour,

Il faudrait que Désiré envoie les plans des pièces à usiner. (bride et pièces de train)

En ce qui me concerne, je n'usine que les pièces de tournage.
Les pièces de fraisage sont à usiner par Adelphe Armange donc il faut passer obligatoirement par Léon Roux.

Bonne soirée

Félicien Wilems

(3)

From: Saturnin To: Félicien

Salut Félicien,

Je t'avais dis que nous pouvions réaliser la bride d'échappement en acier, finalement il serait plus intéressant de la faire en inox (pour améliorer sa durée de vie). Désiré m'a dit que Aries packaging devait avoir des chutes de plaques d'inox que nous pourrions éventuellement récupérer, as tu des contacts privilégiés avec cette entreprise ?.

Hier, j'ai discuté avec **Adelphe** qui **m'a dit que la halle ne disposait pas de fraise pour travailler l'acier**, après discussion avec Désiré, il m'a certifié que des pièces en aciers avaient déjà été fraisées à l'utt pour le Delvion. **Sais tu si ces fraises existent toujours et où elles se trouvent ? (manifestement Adelphe n'est pas au courant ?).**

Bonne journée

Saturnin

(4)

From: Félicien To: Saturnin

Bonjour,

J'ai une réponse de SAM qui possède de la matière 35CD4 .
Il faut passer chercher l'offre à mon bureau et lui demander s'il possède en dia 100 ou 90 de l'inox 316.
Pour les autres pièces, il n'existe que du rond dans lequel on peut tirer des sections rectangulaires.(Par exemple rond dia 50 si SAM en possède)

Effectivement, je dois posséder une fraise 3 dents dia 10 pour l'acier + un tourteau de surfacage dia 80 .

Je vais te fabriquer les bouchons;
Pour le fraisage, il toujours voir avec Léon et Adelphe.

Bonne journée.

Félicien

Cette négociation ponctuelle est représentative de nombre d'autres plus larges en général et impliquant plusieurs acteurs. Tout au long du déroulement du projet, on constate que l'argumentation est importante. Pour soutenir l'annotation sur le plan de l'argumentation, l'outil prévoit une annotation selon le type d'argumentation portée par le corps de l'annotation. Cette typologie est extraite semi-automatiquement du corpus.

2.2. Une fonctionnalité de communication

Dans le cadre d'une médiatisation de l'annotation discursive, il s'agit donc de soutenir la communication d'une part et de typer l'argumentation d'autre part. Afin de soutenir la communication, notre outil propose l'ancrage d'annotation sur un document ou sur une autre annotation. Cette seconde perspective permet en fait de répondre à une annotation déjà déposée sur un document. De même, l'outil prévoit le multi-ancrage de l'annotation, c'est-à-dire un ancrage sur différentes parties du document. Sur la figure 1, trois cadres (encadré) apparaissent au-dessus du corps de l'annotation. Ils correspondent au trois sélections possibles à annoter en même temps. Cette annotation qui est en train d'être rédigée est en fait une réponse à une annotation existante comme le montre la fenêtre de gauche (cercle clair).

L'indexation en type argumentatif est basée sur une étude du corpus et une extraction semi-automatique des termes basée sur la fréquence. Les principaux types sont : accepter, confirmer, décider, définir, demander, faire un point, justifier, refuser, proposer, problème, question, raisons, solution. Cette indexation est à indiquer dans le cadre « type » (nuage).

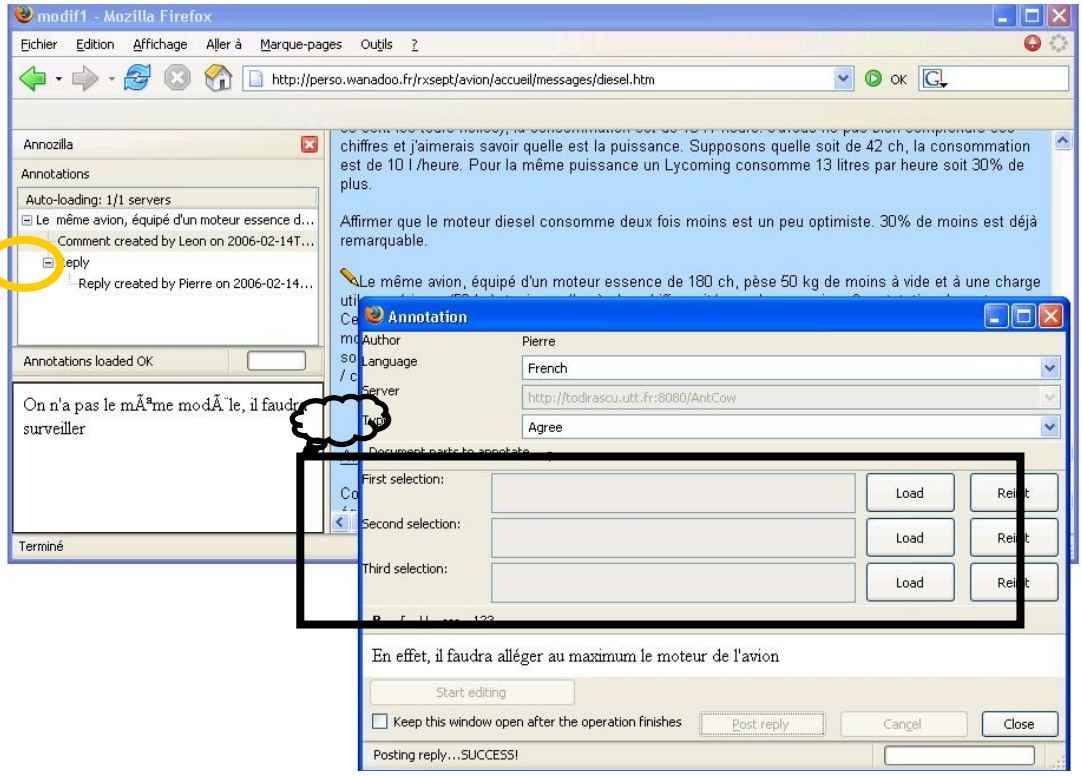


Fig.1 : Réponse à une annotation avec typage argumentatif

Dans le champ des bibliothèques virtuelles, A. Bénel (Bénel *et al.,* 2001) souligne qu'un enjeu est d'offrir non seulement des sources documentaires, mais aussi leur apparat critique. C'est-à-dire un ensemble de lectures liées à la source qui négocient les sens du document et d'où peut naître une étude critique. Cet apparat, les annotations déposées sur un document et les liens vers d'autres documents, est lui-même un document qui peut être critiqué et interprété à son tour. La production de fragments textuels est un processus individuel même dans un cadre de communication. Mais en fixant leurs échanges dans un même document, divers auteurs co-élaborent ce document.

3. Un fragment de document

Dans le champ de l'herméneutique, l'annotation est à part entière un processus d'interprétation du document. L'annotation est un élément riche de l'activité de lecture qui permet l'élaboration de nouveaux documents. A. De Libera (De Libera, 2000) explique que dans l'éducation médiévale, les exercices de lecture sont des exercices de glose consistant en l'ajout de commentaires littéraux sur un texte. Les gloses clarifient une signification d'un terme dans son contexte. Peu à peu, ces notes paraphrastiques deviennent des commentaires organisés. En commentant, le lecteur divise le texte lu en parties déterminées puis l'explique au cours de deux phases principales (la « sententia » et l'« expositio litterae »). Il termine son commentaire par l'examen de plusieurs questions (« quaestio ») reliant le texte à d'autres travaux de référence. Ce type de commentaire est largement lié au texte, mais progressivement, les commentaires deviennent de plus en plus autonomes du texte lui-même. La sententia est alors structurée comme un commentaire à part entière et comme un nouveau type d'argumentation. Elle devient centrale et est à son tour glosée et questionnée. La quaestio, elle, devient une discussion publique, « la disputatio » engageant plusieurs orateurs et rapportée par écrit par un novice. Dans cet exercice, les orateurs argumentent autour d'une sententia (un thème), co-élaborant de nouvelles questions et de nouvelles conclusions. Un autre genre interprétatif naît de la révision de la sententia et de ces questions, c'est la « summa ». La somme est en fait une collection de rapports révisés dans le but d'homogénéiser et d'organiser les connaissances sur un thème. Ces différents types de

commentaires sont issus de lectures et reliés à des documents textuels ou à d'autres fragments textuels. En cela, nous les considérons comme des annotations. De ces descriptions de (De Libera, 2000), plusieurs types d'annotation impliqués dans la création de texte se dégagent :

– la marque physique permettant de diviser un document pour souligner l'importance de certaines parties ;

– la glose qui est un fragment de texte expliquant un terme du document ;

– la note paraphrasant un point du document ;

– le commentaire amenant de nouvelles idées ;

– le commentaire argumentatif qui est une argumentation ou un commentaire organisé construit en coopération par les échanges entre annotateurs.

Dans notre contexte, nous nous limiterons aux annotations textuelles (nous excluons par exemple la marque graphique ou de couleur, ou le croquis), matérialisées sous la forme d'un fragment de texte relié à un document à différents niveaux. Les différents niveaux d'annotations, de la glose au commentaire argumentatif, nous permettent d'identifier des phases d'élaboration de ces fragments textuels. Suite à une lecture, un annotateur peut ancrer une annotation strictement explicative, ou au contraire organiser une argumentation dans le corps de son annotation en vue de l'élaboration collective d'une interprétation sur le fragment saillant.

3.1. Un corpus d'annotation élaborante

Tout au long d'un projet, les participants échangent des méls dans une visée d'élaboration ; élaboration d'un référentiel commun, de solution ou de documents. En étudiant les échanges de mél tels que présentés ci-après, on constate un grand nombre d'échanges qui ont pour but d'élaborer des documents, même encore transitoires. Dans cet échange de (1), (2) et (3), le chef de projet organise une séance de travail avec les bénévoles, et pour cela il doit réussir à mettre en phase des chantiers et des disponibilités de bénévoles. Pour cela, il doit vérifier la faisabilité d'un chantier puis mettre au courant (toujours par mél) tous les membres de l'association. Ces trois méls appartiennent à un échange beaucoup plus long (environ 40 méls tout auteur et destinataire confondus) qui définit la faisabilité des différents chantiers.

Suite à une demande de Saturnin pour relancer le projet en veille depuis le début de l'hiver, Désiré fait une liste des principaux chantiers à mener (1). Chaque chantier dépend de la mise à disposition de pièces et donc de financements. Saturnin le chef de projet doit veiller à la bonne articulation de ces phases. Il lance donc une négociation avec le CA et c'est de la réponse du CA que dépendra la suite du projet. (3) est en fait un résumé récapitulatif de ce qui va être mis en place.

From: Désiré To: Saturnin
salut,
voici les taches que nous devrions pouvoir accomplir les 25/26:
-finition du circuit electrique et essais Il me faudrai la pince a serir les cosses, la demander a Jules
-installation de l'échappement. s'assurer auprès de paquito et félicien que j'aurai la bride pour le 13/02 au plus tard.
-finalisation du circuit de reffroidissement
-finalisation du circuit de fuel
-maquettage de l'orientation du train avant. s'assurer auprès de paquito et félicien que j'aurai les pieces pour le 13/02 au plus tard
-lancement des travaux a effectuer par vous avant les vacances de paques.....

a bientot

d. Louis

(1)

From: Saturnin To: Jules, Félicien

Bonjour,

Désiré a pris contact avec moi, il aurait besoin de la bride de l'échappemment et des pièces du train avant pour le 13/02 au plus tard, est ce envisageable ?. Merci de me répondre au plus tôt.

Bonne journée.
Saturnin

(2)

From : Saturnin To : All

Bonjour à tous,

Voici les tâches qui sont planifiées pour le 25 et 26 février :
- Finition du circuit électrique et essais
- Installation de l'échappement
- Finalisation du circuit de refroidissement
- Finalisation du circuit de fuel
- Maquettage de l'orientation du train avant
- Lancement des travaux à effectuer par les membres impliqués avant les vacances de pâques

Nous restons en attente de vos disponibilités pour organiser ces deux journées.

Bonne journée

Roger et Saturnin (3)

3.2. Une fonctionnalité d'élaboration

Dans le cadre de la médiatisation de l'activité d'élaboration de document via des annotations sur des documents, l'annotation pour la communication est bien sûr centrale, puisque de la négociation naît de nouveaux fragments qui seront la base de nouveaux documents. Au niveau de notre outil, cela va se traduire par une possibilité de collecter des annotations et leur contexte d'ancrage afin de créer un brouillon, base d'un nouveau document. Ainsi, une fois des annotations déposées sur un document ou des documents, il est possible de les choisir selon leurs index (auteur, date, genre argumentatif, …) et de les intégrer dans un brouillon qui est éditable dans l'outil. Ce brouillon est partagé par tous, ce qui implique qu'il est lui-même « annotable » et peut donc donner à son tour naissance à un nouveau document. La figure 2 montre la création d'un brouillon (fenêtre centrale) conservant les traces de communication (annotation de qui, réponse de qui et à quelle annotation). Ces brouillons sont directement reliés à des documents partagés (documents et annotations) ce qui permet de conserver le contexte de production d'un document.

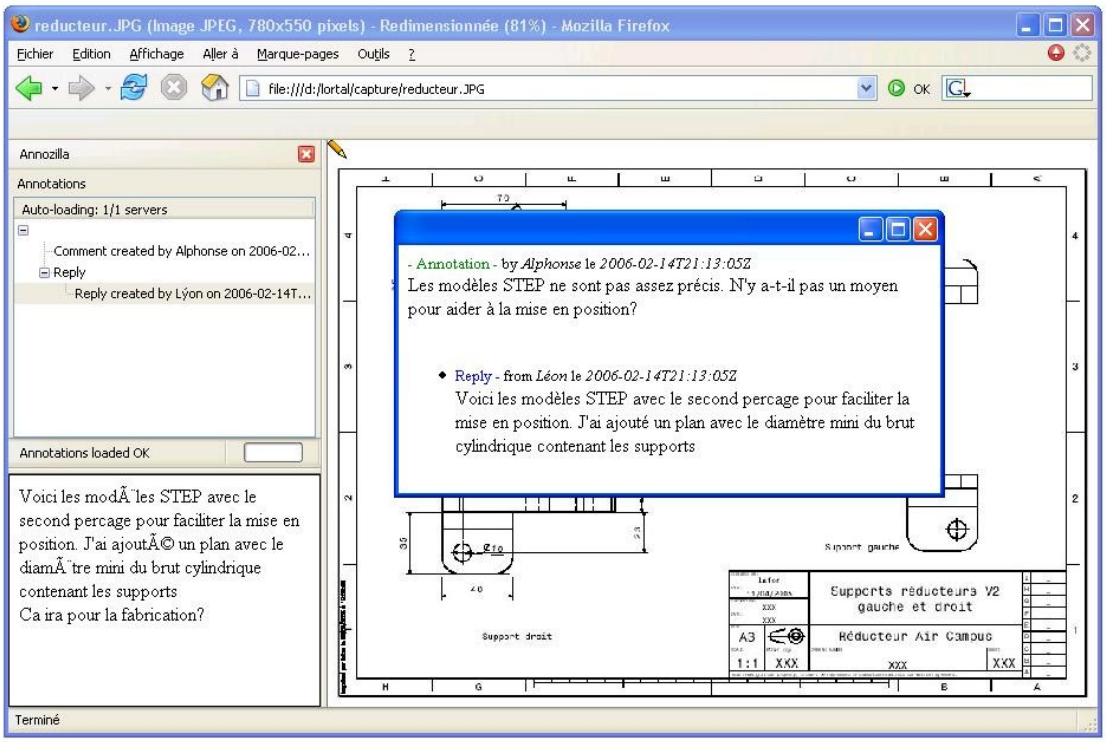


Fig.2 : Création d'un document par collection d'annotations

Tout document est une base potentielle pour l'élaboration d'autres documents, puisqu'il donne lieu à des interprétations qui permettent d'étendre les concepts des uns et des autres. Cependant, pour être partagé, le document doit être disponible et donc répondre à une classification adaptée pour être facilement retrouvé. De même, pour être engagé dans une interprétation ou une argumentation, le document (ou le fragment de document) doit être indexé finement afin de permettre aux différents protagonistes de comprendre son contexte d'interprétation et de tracer la logique de conception du document. Cette classification construite par et pour ce collectif est elle-même un document accessible en élaboration constante.

4. L'annotation pour l'indexation

Médiatiser l'activité d'annotation dans ses différentes dimensions est donc à la fois un moyen de soutenir la discussion autour de textes élaborés ou en cours d'élaboration, mais aussi de concevoir un référentiel commun (Clark, 1996) dans le groupe. Dans les contextes de coopérations émergentes, où les compétences pour la collaboration dans le groupe ne sont pas encore bien établies, les annotations permettent de noter, organiser, explorer, partager des idées, introduire de nouveaux concepts et techniques, créer des alliances ou encore créer une compréhension partagée de certains problèmes. C. Lee parle alors d'artefacts de négociation de frontières (« boundary negotiation artifact » (Lee, 2005). Ces artefacts de négociation de frontières ont un rôle comparable à celui des objets intermédiaires (Boujut et Blanco, 2003) ou des prototypes définis en ingénierie de la conception (Subrahamanian *et al*, 2005). Dans des phases de réflexion « dans l'action » du projet (« in action », Schon et Bennet,1996), les concepteurs conçoivent un objet par une création continue en s'écoutant et en répondant aux surprises de la conversation, en négociant la compréhension mutuelle. L'annotation est le support de cette négociation.

En participant à la construction négociée d'un référentiel commun au groupe, un acteur reflète son engagement dans la volonté de construire une communauté. (Simone et Sarini, 2001) définissent les schèmes de classification (Classification Schemes) comme des objets définis au cours d'une construction collective, et permettant de structurer les connaissances d'un groupe selon une organisation commune. Intégrer un document dans une classification physique (thésaurus), c'est aussi l'intégrer à ses concepts individuels selon un schème de classification qui peut être partagé. Cette intégration de documents dans un schème individuel revient à comprendre un document dans un contexte spécifique. (Simone et Sarini, 2001) soulignent que, malgré l'importance des schèmes de classification dans les mécanismes de coordination, ceux-ci ne sont pas suffisamment observés sur le plan de leur évolution et mise à jour.

Les documents, comme les schèmes de classification évoluent et sont élaborés au fur et à mesure de l'apparition de nouveaux objectifs, de nouveaux membres du projet, de nouveaux documents. Ce sont des documents en élaboration constante, en action, qui impliquent des mouvements endogènes : les documents donnent naissance à de nouveaux concepts qui s'intègrent dans ce même document au cours des versions, et exogènes : les documents font émerger de nouvelles classifications et documents possibles, reliés au contenu des premiers mais extérieurs à ce contenu tout de même ; ce sont des DOcuments Pour l'Action (DoPA) (Zacklad, 2006). On peut soutenir cette génération de DoPAs par le biais d'annotations suivant ces deux grands mouvements issus du principe du cercle herméneutique (Gadamer, 1996) et qui permettent une interprétation des textes toujours renouvelée. En structurant l'ensemble des échanges et des documents produits par le groupe, il serait possible de comprendre les négociations et argumentations intervenues dans la construction d'éléments ou produits du projet. Cette structuration est évidente dans tout groupe qui se constitue, comme le montre l'étude de notre corpus.

4.1. Un corpus d'annotation indexante

La collection de documents numériques est un élément partagé qui permet d'ancrer des tâches distribuées dans un référentiel commun (Clark, 1996) autorisant ainsi une médiatisation de ce travail. Les échanges permettent l'élaboration d'un référentiel commun, qui se traduit par une élaboration collective de schèmes de classification lorsqu'il s'agit d'organiser des documents. Dans un projet, un acteur va ainsi participer à la création de documents et à leur classement selon un plan de classement défini.

Lors d'échanges au cours d'un projet, les utilisateurs vont naturellement vers une indexation de leur document. Cette indexation est particulièrement visible dans la structuration des méls. En effet, lors de l'observation de notre corpus, nous constatons l'utilisation des titres de messages (« subjects » dans la figure 3) pour indexer le message interne voire les pièces jointes au message. Ces index sont utilisés pour retrouver le thème d'un échange, principalement par rapport au domaine auquel a trait le message (« matières brutes, chargeur, brides,… »), mais aussi pour indiquer la phase de planification abordée (« répartition des tâches ») ou encore pour noter quel type d'argumentation porte le message envoyé (« question »).
Il est ici bien clair que cette indexation est modifiée selon l'étape du projet puisque par exemple sur le thème général « Brides et Support Compas » :

a) >Subject: Brides et support compas
b) >Subject: Re: Plans Brides et suport compas
c) >Subject: RE : Delvion : fabrication des brides et du support compas,

la première occurrence est générale, la seconde est focalisée sur la partie « plan », quant à la dernière, elle indique que nous ne sommes plus dans une phase de conception mais dans la phase suivante de « fabrication ». Ce type de classification *ad hoc* montre bien l'importance d'avoir une classification évolutive, non figée et adaptée au domaine.
De même au niveau planification, les exemples :

d) >Subject: Re: Delvion : CR à imputer pour la réalisation
e) >Subject: Re: Imputations sur CR132C,

montrent l'utilisation de titre du message pour classifier au niveau du domaine (« Imputation, CR »), mais aussi pour communiquer des informations (« CR132C) et marquer l'argumentation (« Re » indiquant une réponse). Comme dans toute classification, il est possible de retrouver des « cagibis », des ensembles neutres comme « projet Delvion » ou « Divers Delvion »(fig.3) servant à différencier ces messages de messages hors-projet.

Domaine :

>Subject: DAIR-100
>Subject: Delvion : Dimensions des pièces brutes

>Subject: Delvion : fraise acier
>Subject: Re: Delvion : Pb de fraise

>Subject: Tolérances d'ajustement
>Subject: Delvion : Support Compas
>Subject: Fwd: Matière brute
>Subject: Etude de la commande des gazs

>Subject: Modèles CAO des pièces
>Subject: Poste à soudure TIG
>Subject: Re: Chargeur de batterie

>Subject: Re: Bride d'échangeur
>Subject: Bride d'échangeur
>Subject: Devlion : Bride échangeur
>Subject: Delvion : Bride de l'échappement et pièces
>Subject: Matériau des brides de l'échapement
>Subject: Fwd: Delvion : Bride d'échapement
>Subject: Devlion : Bride et pièce de train
>Subject: Re: Pièces: Bride et train
>Subject: Re: Pièces: dride et train
>Subject: Brides et support compas
>Subject: Re: Plans Brides et suport compas
>Subject: RE : Delvion : fabrication des brides et du support compas

Planification :

>Subject: Re: Adresses non correctes
>Subject: Re: Compléments d'adresse

>Subject: Re: Pièces prévues pour le 13 Février
>Subject: Re: Répartitions des tâches
>Subject: Re: réponse aux questions
>Subject: Re: Tâches à accomplir le 25/26 février
>Subject: week end de février
>Subject: Re: Bilan du Week-end du 25/26 Fevrier

>Subject: Re: Delvion : CR à imputer pour la réalisation
>Subject: Re: Imputations sur CR132C

Argumentation :

>Subject: Question à propos de la bride
>Subject: Question à propos du support compas

Divers :

>Subject: Re: divers DELVION

Fig.3 : Sujet des échanges organisés par points de vue sur le projet

Dans cet ensemble de méls, nous reconnaissons alors bien l'élément joint (la pièce jointe, le texte cité), son annotation (le nouveau message laissé par un auteur), et l'indexation de cet ensemble (le titre, le(s) thème(s) de l'annotation).

4.2. Une fonctionnalité d'indexation pour tracer la collaboration

Afin de soutenir ces perspectives d'indexation et de construction de classification suite à des échanges *via* des annotations sur des documents partagés, nous avons développé une fonctionnalité d'indexation pour notre outil. Les échanges d'annotation sont analysés pour construire une classification commune. L'intérêt d'un corpus tel que le nôtre est qu'il permet de représenter différents points de vue sur le projet. En effet, l'annotation représente à la fois le point de vue d'un acteur du projet (organisationnel), les thèmes abordés par le projet (domaine), les étapes par lesquelles le projet passe (planification) ainsi que les négociations qu'il y a pour l'élaboration de solutions et des documents du projet (argumentation). L'annotation est donc représentative de multiples points de vue sur un objet : organisationnel, de domaine, de planification, d'argumentation.

La classification utile à l'utilisateur doit donc être une classification multi-point de vue évolutive et adaptée au projet. Une telle classification est coûteuse à mettre en place et à mettre à jour, c'est pourquoi notre outil est basé sur un module de Traitement Automatique des Langues qui extrait des candidats-termes pour soutenir l'activité d'indexation des annotations. La classification étant négociée entre les utilisateurs, ce module propose des termes que l'utilisateur accepte ou non et l'outil prévoit l'ajout manuel d'index. Ce module d'indexation basé sur Syntex (Bourigault, 2005) est en cours d'encapsulation dans notre outil.

5. Conclusion

Nous souhaitons soutenir la coopération dans un groupe grâce à l'activité d'annotation. C'est-à-dire soutenir l'annotation dans un but de communication, de création de document ainsi que de création de classification pour organiser les annotations. L'étude de ce corpus sous différents angles nous permet de porter plusieurs regards sur l'annotation et de définir différentes fonctionnalités nécessaires à notre collecticiel pour soutenir une activité collaborative d'annotation. On constate que les annotations sont des éléments polymorphes et complexes permettant aussi bien l'élaboration de documents, que de nouvelles négociations sur les concepts inhérents à ce document et servant à sa classification par une communauté.

Il s'agit désormais de tester empiriquement nos hypothèses sur ces fonctionnalités et d'évaluer dans quelle mesure cet outil permet de médiatiser l'annotation. C'est-à-dire qu'il s'agit de valider les principes et techniques d'ancrage de fragments textuels, mais aussi d'organisation de fragments (classification / indexation) pour la ré-utilisation de ceux-ci pour produire de nouveaux documents. Une expérimentation est mise en place pour observer la constitution de documents en rédaction collaborative.

BIBLIOGRAPHIE

BENEL, A. *et al.* 2001. Truth in the Digital Library: From Ontological to Hermeneutical Systems, *Proceedings of the fifth ECDL*, Lecture Notes in Computer Science, Springer-Verlag Ed., pp. 366-377.

BOUJUT, J.-F., BLANCO, E. 2003. Intermediary Objects as a Means to Foster Co-operation in Engineering Design, *Computer Supported Cooperative Work* 12(2), pp. 205-219.

BOURIGAULT, D. *et al.* 2005. Syntex, analyseur syntaxique de corpus, in *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles,* Dourdan, France.

CLARK, H. 1996. *Using Language*, Cambridge University Press, Cambridge.

DE LIBERA, A. 2000. *La philosophie médiévale*, Paris, PUF (« Que sais-je ? » 1044), 4e éd.

GADAMER, H-G. 1996. *Vérité et méthode. Les grandes lignes d'une herméneutique philosophique* (1960) ; trad. Fruchon, Grondin, Merlo, Seuil, 1996 ; Vol.1 des Gesammelt Werke, Mohr, Tübingen, 1986.

HAROCHE-BOUZINAC, G. 2000. Une esthétique de la brièveté, *Revue de l'AIRE – Recherches sur l'épistolaire*, n° 25-26, pp. 49-51.

LABBE, H. & MARCOCCIA, M. 2005. Communication numérique et continuité des genres : l'exemple du courrier électronique, *Texto* ! (http://www.revue-texto.net/Inedits/Labbe-Marcoccia.html)
LEE, C. 2005. Between Chaos and Routine: Boundary Negotiating Artifacts in Collaboration, *Proceedings of the Ninth European Conference on Computer Supported Cooperative Work,* pp. 387-406.
MARCOCCIA, M. 2001. L'animation d'un espace numérique de discussion : l'exemple des forums usenet, *Document Numérique*, Vol. 5, n°3-4, pp.11-26.
MARCOCCIA, M. 2004. On-line Polylogues : conversation structure and participation framework in Internet Newsgroups, *Journal of Pragmatics*, vol.36 n°1, pp.115-145.
SCHON, D., BENNET, J. 1996. Reflective Conversation with Materials, in T. Winograd (éd), *Bringing Design to Software*, Stanford University and Interval Research Corporation Addison-Wesley, 1996 chp.9 http://hci.stanford.edu/bds/9-schon.html
SIMONE, C. and SARINI, M. 2001. Adaptability of Classification Schemes in Cooperation: What does it mean?, *Proceedings of ECSCW 2001*, pp. 19-38.
SUBRAHMANIAN, E., MONARCH, I., KONDA, S., GRANGER, H., MILLIKEN, R., WESTERBERG, A. 2003. The N-Dim Group, Boundary Objects and Prototypes at the Interfaces of Engineering Design, *CSCW Journal*, Volume 12 , pp. 185-203.
ZACKLAD, M. (à paraître) Documentarization processes in Documents for Action (DofA): the status of annotations and associated cooperation technologies, in *JCSCW.*

L'ÉDITION EN LIGNE AUJOURD'HUI
SELON QUEL MODÈLE ÉCONOMIQUE ?
Balisage des textes numérisés.

Constance KREBS
Conseil éditorial et doctorante Paris 3/ Censier-Sorbonne Nouvelle

Résumé : *Les imprimeurs ont établi, dès l'apparition de l'informatique, des balises qui permettent de créer et de développer à l'infini, tout en le structurant clairement, un protocole de mise en page. Les éditeurs qui ont réfléchi sur la numérisation d'un texte ont, eux aussi, élaboré une TEI (Text Encoding Initiative) qui correspond à l'encodage d'un texte, autrement dit à l'élaboration de sa structure (ou mise en page) à l'aide de balises.*

Cette TEI, qui prévalait sous SGML, était capable de rendre le texte lisible, la mise en page conservée sur tout support de lecture (qu'il soit électronique ou traditionnel, Cf l'expérience princeps de balisage SGML chez l'éditeur 00h00 : papier, écran d'ordinateur, e-Book, PDA…). SGML est remplacé par l'XML, on évite donc la TEI tout en permettant une structuration du texte beaucoup plus large, et plus fiable. En effet, grâce à une arborescence lisible, l'XML permet de structurer un texte littéraire, ou scientifique, mais aussi de l'enrichir par des « valeurs ».

Dès lors, les textes – qu'ils soient littéraires ou scientifiques, documentaires ou poétiques, écrits en toutes lettres ou en notes de musique – peuvent se lire, et s'étudier, selon un encodage établi par l'éditeur. Cet encodage que l'éditeur établit manuellement, à l'aide d'un protocole de balises extrêmement précis, offre au lecteur un outil d'investigation.

Ce système de valeurs est établi à travers les feuilles de style que tout éditeur utilise. Or, lier au gabarit de mise en page un protocole qui attribue un système de valeurs rend robuste, obligatoire et transparent, presque indolore financièrement pour les éditeurs, tout ce travail préparatoire à la numérisation et à la recherche par champs.

Ainsi, pour les ouvrages littéraires, scientifiques, les documents ou les guides touristiques, allons-nous établir des automatismes d'indexation. Que l'on indexe, que l'on retrouve le mètre calculé au plus juste lorsqu'il s'agit de poèmes, que l'on compare les mètres du texte ou dans différentes œuvres, tout cela est désormais possible

On peut encoder à l'infini un texte à condition que cet encodage soit conduit selon une arborescence et une structure extrêmement logique. Et pour des raisons pratiques que l'artiste ignore, l'éditeur ne peut pas se permettre d'établir un encodage différent pour chaque texte – aussi doit-on mettre au point un protocole de balisage qui soit le plus fin possible.

Ce projet d'édition de textes balisés permet d'éditer des ouvrages classiques pour un public scolaire, de mettre à la portée des étudiants des outils automatiques abordables pour l'étude des textes littéraires, les partitions musicales, les récits hypertextes et tout texte structuré. Il s'agit d'un modèle économique, nullement définitif, d'édition en ligne d'ouvrages de littérature générale.

Les éditeurs numérisent des textes, qu'il s'agisse de nouveautés ou du fonds. Mais pour quelles raisons numérise-t-on ? Que va-t-on faire de ces textes numérisés ? À quel public sont-ils destinés ? et selon quels protocoles va-t-on numériser ?

Pourquoi numérise-t-on puisque le texte brut mis en ligne ne rapporte rien à l'éditeur ? Il n'envisage pas d'apport éditorial. C'est seulement un outil promotionnel, simple outil de communication de la société des écrans. Du coup, il risque de s'épuiser en même temps que les librairies traditionnelles fermeront.

Concurrence des supports

Si le texte mis en ligne est brut, sans valeur ajoutée, cela sert Google, Amazon, mais cela dessert la maison d'édition – et le métier d'éditeur. En Grande-Bretagne, Google couple le moteur de recherche, les extraits de textes des livres disponibles sur le marché avec une librairie. Via Google Map, un plan donne accès à la librairie la plus proche du village ou du quartier, en fonction de l'adresse de l'internaute. Mais la qualité de lecture du texte extrait du livre est toujours assez déplorable. En France, les éditeurs se sont pour le moment opposés à cette fonction de Google.

Or paradoxalement, la perte de qualité du livre ou des extraits mis en ligne dessert le métier d'éditeur, et contribue à léser la librairie traditionnelle. En France, la vente de livres en ligne sans valeur ajoutée, tels quels, comme on le voit aujourd'hui dans les librairies en ligne ou quelques

sites qui proposent des e-books ou livres numériques, ne représente rien de neuf par rapport au papier. Lecture inconfortable, odeur absente du papier, entend-t-on, je n'y reviens pas. Bref, c'est invendable, même sur supports dédiés (palm, lecteur électronique, papier électronique) hors de prix (ou seulement à des marchés de niches...). La solution réside dans la gratuité du support ou du contenu.

Support de la concurrence

En effet, le client ne voit pas encore (dans 50 ans peut-être) l'intérêt *d'acheter* un texte brut en ligne – il ne voit même pas l'intérêt d'un texte brut en ligne. La vidéo, le son, la documentation ont leurs usagers sur le Net, car ils sont plus dynamiques… Mais si ce livre numérique, qu'on appelle à tort aujourd'hui *e-book*, est disponible *gratuitement* en ligne, c'est différent. Il devient alors, réellement, un outil de communication. Le Net est du coup, et seulement, une base qui informe mais aussi qui encourage l'achat en librairie. C'est vrai pour la philosophie, les essais, toute science humaine – mais c'est aussi vrai pour le roman, la poésie, le théâtre, toute littérature générale. Pour que cette ouverture de l'édition vers la toile soit néanmoins valable, le livre numérique doit être fabriqué par les éditeurs eux-mêmes - ou selon des critères qu'ils ont eux-mêmes définis. Car c'est ainsi que la mise en ligne du livre pourra correspondre aux attentes de l'éditeur.

C'est ce qu'ont très bien compris, et depuis longtemps, Patrick Cahuzac (Inventaire/ Invention), Michel Valensi (l'Eclat) et Paul Otchakovski-Laurens (POL). Cela dit, ni Inventaire/Invention, ni les Editions de l'Eclat, ni POL ne proposent un quelconque enrichissement des textes (ou des dessins) gratuitement mis en ligne. Rien de dynamique, rien de *clicable*. Le lecteur ne peut qu'opter pour deux options : imprimer, lire, jeter les feuilles, ou bien enregistrer le fichier après lecture. Mais comme le lecteur n'aime la lecture à l'écran que s'il joue avec sa souris, en général il parcourt le fichier avant d'acheter l'ouvrage imprimé. C'est cette troisième façon de lire qui permet à Inventaire/ Invention, à l'Eclat de survivre, à POL de vivre mieux.

Vers la complémentarité : la gratuité est payante

Cette gratuité du fichier numérique encourage l'achat en librairie. Cet échange entre Net et commerce traditionnel apporte un soutien considérable aux petits et moyens éditeurs qui ont accès à un immense marché. À cet égard, l'exemple des Editions de L'Eclat est éclairant. Michel Valensi a réussi le tour de force de diffuser aussi bien en ligne qu'en librairie.

Les résultats d'une enquête (élaborée en février 2006 par TNS/ Sofres) qui sont parus dans Livres Hebdo 637 le 17 mars 2006 vont éclairer encore mes propos. 19% des personnes interrogées achètent un livre après une recherche sur le Net. Cependant, parmi les lecteurs 32% ont entre 25 et 49 ans, mais 7% sont âgés de 20 à 24 ans, et 6% de 15 à 19 ans. Or, ces 11% de lecteurs ne consultent pratiquement qu'en ligne (LH ne donne pas les chiffres). On peut envisager que, dans un proche avenir, la totalité de ces gens, soit 45%, iront s'informer directement sur les réseaux numériques. Mieux, ce seront à travers les réseaux numériques qu'ils reviendront à la lecture. Aussi, les librairies doivent être soutenues, certes, mais pas selon les usages d'aujourd'hui. Continuons d'explorer ce retournement de situation qu'implique la révolution numérique.

L'internaute lit… sur Internet

Les éditeurs ont longtemps eu peur du Net. Désormais, nous devons compter sur le Net pour rester éditeur. Pourquoi ? Si l'on regarde les chiffres de Médiamétrie de la fin de 2005, on constate que :

46 % des foyers français sont équipés d'un ordinateur (contre 38,8 % fin 2002) en avril 2005, contre 50,6% en avril 2006, (soit une progression d'environ 4% par an)

71 % de foyers avec enfants sont équipés d'un ordinateur au 2e trim 2005 (contre 69% trois mois avant),

plus de 50 % de Français sont internautes (35% fin 2002) (même s'ils ne disposent pas d'ordinateur chez eux), soit en moyenne 5% d'augmentation par an ; sur cette base une projection en 2010 donne 75% de la population comme internaute.

52 % de foyers avec enfants sont équipés d'internet au 2e trim 2005 (contre 47% trois mois avant),

69 % des foyers sont connectés à haut débit en avril 2005, 85% en avril 2006. Soit une augmentation de 41% en un an.

57 % des internautes achètent en ligne (contre 23 % d'entre eux fin 2001).

Par conséquent, les internautes achètent en ligne de façon exponentielle par rapport au nombre de foyers connectés et consultent de plus en plus vite grâce au haut débit. Il est intéressant de remarquer que, selon la délégation Internet du ministère de l'Education nationale, l'école participe à hauteur de 68 % dans la formation des enfants de plus de 11 ans à l'usage d'Internet, que 75 % d'entre eux utilisent l'ordinateur de leur foyer pour préparer un travail scolaire (contre 69% des étudiants, classe d'âge plus élevée qui, pourtant, est équipée à 100% d'internet). En classe, dans les collèges et lycées 40 % des profs utilisent le Net – et 56% des parents pour aider leurs enfants. L'ordinateur est désormais l'instrument incontournable et intégré pour la recherche de documents : selon le ministère de l'Education nationale, si 31% des 19-24 ans disent ne jamais utiliser l'ordinateur dans leur activité universitaire, ils ne sont plus que 2% parmi les élèves de 11-18 ans (http://delegation.internet.gouv.fr/barometre/index.htm).
Enfants, nous n'avions que les livres, la radio et la télévision pour nous documenter. Aujourd'hui, la première source d'information des 15-25 ans est le Net. Blogs, MSN, infos, c'est la recherche de l'interaction qui fait l'internaute. Si l'éditeur ne met pas ses livres en ligne, ou sur supports électroniques dédiés à la lecture, il n'a plus de raison d'être.

L'éditeur publie sur papier…
Le seul intérêt d'une mise en ligne basique de textes réside précisément là. C'est une plus grande diffusion – mais pas seulement. Le texte mis en ligne doit être gratuit et sa visibilité relayée par des partenaires. Valensi vend en librairie traditionnelle 1% des livres dont les « lyber » ont été lus (partiellement ou totalement) en ligne. Avec un référencement Google, soit 60 000 livres numériques lus depuis six mois, cela donne pour lui 600 exemplaires vendus en plus, tous ouvrages confondus. Sur un catalogue de petit éditeur qui, même bien diffusé, arrive à vendre au maximum 50 exemplaires par an d'un ouvrage, c'est énorme. Mais cela ne suffit pas.
Certes, il faut absolument mettre en ligne, mais si la mise en ligne est une chance pour l'édition traditionnelle, comme le soutient « Le Monde » dans un important dossier régulièrement mis à jour, elle peut représenter un danger comme le défendent les auteurs de la SGDL.
La position d'éditeur n'a plus de sens sur le web à l'heure où tout auteur peut mettre son manuscrit en ligne, peut écrire son blog aussitôt édité, diffusé et référencé. On trouve par exemple d'excellents sites d'auteurs qui sont des œuvres à part entière, et dont le trafic approche les 5 000 visiteurs par jour à raison de plus d'une minute par page : www.tumulte.net, www.désordre.net, www.le-terrier.net. Il faut investir ce qu'on considère comme un marché de niche – car ce qui était vrai pour le papier ne l'est pas forcément pour le numérique. D'autres auteurs (parfois les mêmes) se sont regroupés pour faire vivre leurs textes les plus difficiles sans l'aide d'une revue papier qui plafonnerait à 300 exemplaires par trimestre. www.fluctuat.net, www.remue.net, www.chaoïd.net, http://lafemelledurequin.free.fr, etc. La revue de création littéraire *Remue* approchait en 2005 les 100 000 visiteurs par mois et elle atteint aujourd'hui près de 160 000 visiteurs par mois… Quand on sait qu'une revue peine à dépasser les 300 exemplaires vendus et qu'un romancier confirmé atteint rarement les 3 000 exemplaires vendus, on reste pantois devant de tels chiffres. Par conséquent, si ces revues en ligne ont leur équivalent en librairie, en version papier ou sur support numérique, je pense qu'elles deviennent rentables.

… et il édite en ligne
Alors comment conserver ses compétences d'éditeur et améliorer ses résultats, sa visibilité, en mettant des textes en ligne ? En faisant ce qu'un éditeur a toujours fait : enrichir un texte. Si cet enrichissement est un peu différent, l'intervention de l'éditeur lui redonne toute sa place, sa légitimité.
Que va-t-on faire pour conserver notre rôle ? Au système de balises de mise en page classique, on ajoute un système de balises qui enrichit le texte. L'éditeur peut donner du sens à l'œuvre et des directions au lecteur.
Sur le Net, selon Médiamétrie, 80 % des documents recherchés sont à contenus éducatifs, culturels ou scientifiques, et l'ordinateur sert essentiellement à la recherche documentaire (80%), au traitement des textes (71%) et au courrier électronique (58 %). Le net est donc un outil de documentation. Que ce soit pour la médecine, l'information, les Sciences humaines et sociales, la littérature. Par conséquent, à l'heure actuelle, pour inciter le lecteur à lire sur écran (ou à avouer qu'il lit sur écran, parce que le Net c'est comme la télévision, tout le monde l'utilise mais personne

ne l'avoue), l'éditeur doit ajouter une valeur documentaire au texte - et cet outil de documentation prend deux directions.

Atours et détours

La première direction permet au lecteur internaute de se documenter grâce aux très grandes ressources qui sont en ligne. Cela est d'une grande importance pour l'éditeur. On peut imaginer un site dédié à un best-seller, avec des pages ressources, des jeux, des commentaires, liens, chats avec l'auteur, etc. L'essentiel est que nous proposions des interactions à partir du livre – à une époque où les adolescents écrivent tous sur leurs blogs.

La seconde direction, qui est complémentaire, permet l'autodocumentation d'un texte. Par un système de balisage sémantique du texte, l'éditeur peut créer :

- des index automatiques (lieux, personnes, mots-clés, notions, couleurs, thèmes, etc.) qui ouvrent vers quelques sites et quelques œuvres à découvrir,
- pour les enseignants et leurs élèves ou étudiants, des balises morphologiques qui indiquent la forme d'un texte littéraire (métriques du vers, didascalies du théâtre et autres rythmes, grammaire, syntaxe, etc.[1]),
- des liens de chaque mot vers sa définition pour des recherches lexicales[2],
- les occurrences des mots, comparaison possible avec d'autres œuvres de la même époque, du même auteur, par une représentation sous forme de cartes, par exemple.

Tous ces points ne sont que des exemples, à développer selon les collections et les besoins. Ces systèmes de balises sont à la fois des outils pour interroger une œuvre, et des illustrateurs pour rendre le livre numérique ludique et vivant de sorte que la *relation au livre soit interactive*.

Retour d'expérience

L'expérience d'édition en ligne prouve que la seule impression à la demande liée à un fichier *payant* en ligne ne suffit pas à faire vivre une maison d'édition. En 1997, lorsque Jean-Pierre Arbon et Bruno de Sá Moreira ont commencé à réfléchir à la mise en ligne de textes pour ce qui allait devenir 00h00.com, ils m'ont confié le choix et le protocole de mise en page des textes, ainsi que leur relecture après OCR. Le catalogue dans lequel j'avais à choisir était celui d'œuvres classiques. Nous travaillions en SGML, de sorte à faire un PDF lisible sur des supports dont nous ne savions pas encore vraiment ce qu'ils allaient être. Fin 1998, nous nous sommes intéressés au Palm Pilot, et Olivier Pujol est venu nous présenter une maquette de ce qui allait devenir le Cybook, pendant que Franklin fabriquait son Rocket eBook avant de le revendre à Gemstar. Au même moment, chez 00h00, chacun relisait et mettait en ligne un ouvrage par semaine, de sorte qu'il y ait environ 15 livres disponibles par mois. Théâtre, poésie, roman, essai, chaque genre littéraire avait un protocole de mise en page particulier. Car, à chaque mise en page, j'avais imaginé des feuilles de style qui auraient dû permettre aux étudiants et aux lycéens, ainsi qu'au grand public curieux, comme on dit sur les quatrièmes de couverture, d'interroger depuis le site de la maison d'édition, chacun des ouvrages que le lecteur aurait eu à sa disposition. Celui-ci disposait d'un salon, dans lequel il entrait à l'aide de son mot de passe, avec forum dédié, et forum général. Ce salon aurait dû être sa bibliothèque virtuelle, sorte de club où se seraient retrouvés des lecteurs élus par des affinités communes.

Mais le salon n'était pas aménagé.

En effet, il nous a manqué un élément, de taille. Nous avons évité de nous interroger sur les besoins du lecteur. Ou plus exactement, parce qu'il était encore trop tôt pour que le lecteur

[1] Si dans *Les Châtiments*, on choisit le terme « alexandrin », ou si l'on clique sur ce mètre depuis le PDF codé, on peut retrouver tous les alexandrins du recueil. Si l'on répète la même requête dans l'œuvre poétique de Victor Hugo, on retrouve les alexandrins, et leur évolution de 1814 à 1884 – ou, du moins, les titres des œuvres dans lesquelles on peut les trouver. Si, à « alexandrin », on relie « théâtre » et « Victor Hugo », on trouve *Hernani*, et les autres pièces. Si l'on demande « alexandrin » et « XIXe siècle », on a les vers romantiques en douze syllabes. Si l'on demande « alexandrin », « XIXe siècle » et « XVIIe siècle », on peut constater à quel point le romantisme a changé le mètre classique.
Et l'on peut en faire un corpus de thèse, un cours de français, une prépa au bac, etc. L'enseignant peut montrer cela sur grand écran en classe. Le curser se mue en main ludique pour montrer les éléments clicables. On peut imaginer des lectures collectives, débutées en classe, qui déboucheraient vers des lectures intimes, sur e-book, e-paper ou PDA. Tout cela ne demande qu'à être développé.

[2] http://www.inlibroveritas.net/lire/oeuvre4606-page1.html

éprouve ce type de besoins, nous avons oublié de nous demander comment le lecteur viendrait au numérique. Pour cela, il fallait un directeur de collection qui connaisse aussi bien l'enseignement du français (en littérature générale) que les possibilités qu'offre Internet. Nous ne l'avons pas trouvé à temps. Le projet est resté en l'état. Les balises sont bien présentes, mais n'ont jamais été exploitées.
Mais comment faire en sorte que le lecteur *lambda* ait envie de chercher le mot « rêve » ? Et d'autres ? Il faut l'entraîner dans la danse des mots... C'est-à-dire, en sus de ce qui vient d'être dit, faire préparer des notices et préfaces dynamiques, des appareils critiques d'un nouveau genre qui excitent la curiosité en donnant des mots-clés comme pistes de lecture, ou pour découvrir quelque chose derrière la lecture, le fameux feuilleté, une sorte de jeu de piste à base de noms propres, couleurs, parfums, objets, etc. (selon le livre, bien évidemment, chacun ayant sa spécificité...). C'est un fait. Mais nous devons communiquer – évidemment. Pour cela nous pouvons mettre en ligne et rendre dynamique les dossiers de presse – qui, du coup, seront à l'usage de la presse et des lecteurs. Dynamiques et ouverts, ils ressembleront à des mini-sites de découverte, avec des liens, du son, des images animées, clips vidéo, etc.

Deux collections, une synergie
Pour cela, envisageons deux collections de base qui, à terme, se rejoindront naturellement par complémentarité. Une collection d'ouvrages littéraires classiques adaptés aux lycéens, avec livre papier et PDF enrichi. Une collection de livres papiers issus d'œuvres numériques, ou mises en ligne qui piqueraient la curiosité des clients de librairies de premier niveau.
Car qui irait chercher les outils du moteur de recherche ? Il faut travailler avec les enseignants, comme le font déjà la plupart des éditeurs de livres scolaires, qui proposent les corrigés, plans de cours, et tout dossier pédagogique, en ligne, dans un espace réservé sur le site de la maison d'édition. Gallimard a ouvert le sien en 2003, en relais de son Cercle de l'enseignement, www.cercle-enseignement.com (revue du Cercle plus étoffée que dans sa version papier, dossiers thématiques, critiques, extraits de livres audio, etc.), Magnard prépare un club pour 2007, Nathan lance son site dédié à la rentrée, ainsi que Larousse et Bordas www.universdeslettres.com, à ce jour le plus interactif (notes et réponses aux questions sur l'image dynamiques, en PHP).

En pratique
Supposons que le client achète un livre en librairie (en ligne ou traditionnelle). Il ne s'intéresse pas au web, cela ne le concerne pas. Nous, qui savons que nous n'allons pas survivre si le lecteur ne se met pas au numérique, qu'avons-nous à proposer ? Dans chaque exemplaire du livre, on glisse un code, un mot de passe qui, couplé à l'adresse e-mail du lecteur, donnera accès à une base de recherche ainsi liée au livre.
Le PDF (gratuit) contiendrait un (ou plusieurs) lien(s) vers une page qui proposerait une maquette du moteur de recherche, expliquant en quoi l'acquisition de cet outil est indispensable. Ce lien, ou son lien retour, proposerait au lecteur-internaute un mot de passe lié au PDF (qui reste à étudier). Une fois les professeurs séduits par l'objet, ce sont les étudiants et les lycéens qu'il faudra achalander.
Par son mot de passe, le lecteur a accès à la recherche. Il range ses données dans une fenêtre dédiée, qui lui est personnelle, un espace lecteur qui est une sorte de salon de lecture virtuel, et privé. Là sont rangés ses PDF, ses requêtes et leurs réponses[1]. Sa base de données personnelle. Offrons-lui aussi la possibilité de prendre des notes, comme le propose le dernier PDF Reader, et de faire, dans une certaine mesure, du copier-coller. Pour le travail, pour les échanges avec les amis, etc. Essayons de déverrouiller le livre numérique, ce qui plaît au lecteur. Ainsi aurait-il le droit de griffonner à l'écran, de copier des citations dans une dissertation, des passages dans un carnet, dans un mail, etc.

[1] Par exemple, si le lecteur choisit des mots comme « liberté », « mort », « enfant », « courage » dans *les Châtiments* de Victor Hugo, il peut, grâce à la base, les retrouver dans les pages 43, 156 et 321 du PDF (qui correspondent exactement à celles du livre papier). Si le lecteur choisit ces mêmes mots pour une requête qui couvre des livres d'Hugo qu'il n'a pas achetés, la base dévoile les titres des œuvres du catalogue dans lesquelles le lecteur pourra trouver ce mot, mais sans les références exactes. Pour avoir les folios et les extraits, il faudra le code d'accès vendu avec le livre. L'éditeur proposera, par exemple, un parcours de formation intellectuelle ou de citoyenneté, éventuellement en relation avec citations d'autres auteurs (contemporains) qui parlent de ces mêmes thèmes (BHL, Houellebecq, Villepin, Kertesz, etc.).

Le PDF gratuit fait la force des Editions de l'Eclat, de Google, d'Inventaire/Invention, de toute la documentation sur Internet. Cela permettra la notion de partage qui viendra, comme le Web 2.0, dans un deuxième temps. L'ouvrage sera téléchargeable, en accès libre – ce qui facilite le référencement des œuvres et leur visibilité, ainsi que leur diffusion. Seul le code d'accès au moteur de recherche serait payant. Offert avec le livre acheté en librairie, il serait payant en ligne, depuis le site de la maison d'édition.
Une fois que l'éditeur a ajouté une valeur au texte numérique, qu'il a ainsi mis en place ce que Google n'offre pas, il peut développer un pôle d'*attrait à la lecture en ligne* en fonction des collections, des publics visés, etc. L'éditeur offre ainsi une œuvre dans toute sa densité, son épaisseur, et sa dynamique.
L'écrivain donne à lire une œuvre. Sa dématérialisation même permet d'en partager des lectures diverses. Les parcours et les cheminements proposés aux lecteurs redonne à l'œuvre toute sa chair. Notre rôle est, peut-être, de révéler que le corps du texte est accessible à tous. L'éditeur met à disposition de chacun une œuvre, des niveaux de lectures possibles.
En outre, le lecteur a intérêt à acheter dans la même collection l'ouvrage imprimé pour avoir accès aux outils disponibles en ligne par le mot de passe. Grâce à la base de données personnelle du lecteur, l'éditeur peut évaluer les centres d'intérêt du lecteur. Ainsi peut-il ensuite envisager un partage des connaissances qui l'aiderait à étendre sa clientèle et qui s'adapterait avec le réseau au web 2.0. Lié à la possibilité de pister les requêtes du lecteur, il aiderait à repérer également les parcours, les thèmes porteurs et rendrait les notices / préfaces à énigmes plus performantes dans les éditions postérieures.
En effet chaque recherche part d'un ouvrage pour aller vers d'autres. Si l'utilisateur U1, qui a lu et recherché dans le livre *a*, ouvre son champ de recherches vers un autre livre, livre *b* – ce qui est commun à tous les lecteurs étudiants. Il rentre dans sa base de données personnelle, ses résultats (notés RU1). Ces résultats associés dans la base à d'autres livres, vont l'orienter vers les livres *c, d, e, f,* qui regroupent les mêmes réponses. Par ailleurs, si un lecteur U2 d'un livre *z*, *y* ou *x*, pose des questions similaires à celles d'U1, la base orientera le lecteur U2 à acheter les ouvrages *a, b, c* et *d* que le lecteur U1 a déjà interrogés dans le même sens. Autre moyen de captiver un lecteur. C'est le principe même du web 2.0 : on voit ce qui marche et on développe.
Ainsi, grâce à ces outils, l'éditeur resterait éditeur de son fonds et de ses nouveautés. On peut aussi envisager un modèle de montée de la visibilité de l'éditeur. Il existe en effet une puissante et rapide communauté internet qui relaie les « bonnes » informations (institutions, presse, blogs, bibliodoc, etc.) et qui dirige en 24 heures plusieurs dizaines de milliers de clics sur un site. Cela nécessite une bonne connaissance de la communauté. Editeur de littérature générale, il deviendrait aisément un éditeur à la fois savant et vulgarisateur. Le libraire resterait dans le réseau de distribution du livre, voire il participerait à son expansion. Les lycéens retrouveraient un intérêt pour les lettres via l'écran qui fascine – tandis que les lecteurs découvriraient des œuvres nouvelles et lisibles sur tous les supports, même le papier !

BIBLIOGRAPHIE

MALLARMÉ, S. *Vers et Prose*, Éditions 00h00.com
HUGO, V. *Les Châtiments*, Éditions 00h00.com
LA ROCHEFOUCAULD, F. de, *Les Maximes*, Éditions 00h00.com
RIMBAUD, A. *Les Illuminations*, Éditions 00h00.com
RIMBAUD, A. *Une saison en enfer*, Éditions 00h00.com
ROUSSEAU, É. *Les Confessions*, PCN/Prototype2PCNConfessions, inédit
LAUTRÉAMONT, Isidore Ducasse dit, *Chants de Maldoror*, Éditions 00h00.com
ROSTAND, E. *Cyrano de Bergerac,* Éditions 00h00.com

LE CORPUS CONÇU COMME UNE BOULE

Étienne BRUNET
Université de Nice

SOMMAIRE

La notion de corpus semble avoir un contour précis et fixe par quoi on l'oppose à d'autres notions plus incertaines, comme le genre, le discours, la langue. Il n'est pas toujours possible de décider qu'un texte appartient à un genre ou qu'un mot appartient à la langue. Mais quand un corpus est constitué, on sait sans contestation si un mot s'y trouve ou non et si un texte y est ou non incorporé. Pourtant, à la réflexion, quand on fixe son attention assez longtemps sur cet objet dur qu'est le corpus, la vue se brouille, l'objet fond et se ramollit comme les montres de Dali.
Tout d'abord un corpus est toujours artificiel. La nature n'en produit pas spontanément. C'est une création nécessairement subjective. Pire encore, la création est orientée, conditionnée par une hypothèse, par un objectif de recherche. Quelques précautions qu'on prenne pour affiner les critères de sélection, pour les justifier et pour les appliquer, il y a toujours des choix à décider, des doutes à faire taire, des contraintes à respecter, des compromis à négocier, un ordre à établir, un terminus *a quo*, un autre *ad quem* à délimiter. Et comme l'opération de traitement est plus rapide que celle de sélection, la tentation est grande de modifier la composition du corpus, au vu des résultats d'un premier traitement, en mêlant ainsi inextricablement et illégitimement les procédures subjectives de la sélection et les procédures objectives du traitement. En tailladant à propos dans le corpus comme un sculpteur dans la pierre, on finit par lui donner la forme souhaitée, compatible avec l'hypothèse initiale et dotée de la fausse garantie d'un traitement impersonnel.
Supposons cependant qu'un corpus ait réuni tous les suffrages, qu'une commission *ad hoc* ait établi les critères, qu'une autre commission indépendante ait procédé à leur application, et que tous les recours ou appels aient été épuisés. Acceptons l'idée d'un corpus à la pureté eucharistique. Reste à l'introduire dans la salle blanche du traitement. Mais ici de nouveau le pas est suspendu devant le seuil à franchir. Il y a plusieurs salles d'opération, plusieurs technologies disponibles, plusieurs logiciels à utiliser et un choix préliminaire à faire : faut-il soumettre le corpus à un traitement purement documentaire, ou à un traitement linguistique ou à un traitement statistique ? Ces distinctions n'ont pas de barrière fixe : beaucoup de logiciels proposent à la fois des fonctions documentaires et statistiques et certaines fonctions à objectif linguistique, comme la lemmatisation, peuvent emprunter la voie statistique.
Supposons le pas franchi et le logiciel idoine. On entre alors dans un labyrinthe. Perplexité devant le trousseau de clés qui s'affiche à l'entrée. On a souvent l'impression de pénétrer dans un jeu de rôles. Heureux l'expert qui sait utiliser la panoplie des outils, et résoudre le rébut des résultats. Passe encore pour les fonctions documentaires qu'un néophyte peut maîtriser sans grand effort. Mais dès qu'intervient la statistique, le non-initié reste perplexe devant les options à choisir, les traitements à opérer, les tableaux à constituer, les graphes à commenter. C'est pire encore lorsque le doute l'épargne et qu'il se promène sans vertige sur le parapet de l'interprétation. L'évidence visuelle dont se prévaut un simple histogramme peut cacher des pièges et des incertitudes : s'agit-il de fréquences absolues, ou relatives ou réduites ? Quelle méthode a-t-on utilisée pour le calcul de l'écart, la loi normale, la loi hypergéométrique ou quelque autre ? Et sur quelle référence, interne ou externe, s'appuie-t-on ? Il arrive parfois que le seul expert apte à débrouiller les fils soit l'auteur du logiciel utilisé. Et comme j'ai cet avantage pour le logiciel Hyperbase, on me permettra

d'en profiter. Qu'on se rassure : je n'abuserai pas de la situation pour décrire en détail le fonctionnement du logiciel. Je m'en tiendrai à une seule fonction : celle qui mesure la distance intertextuelle.
Dans une première approche Hyperbase suit la méthode Jaccard qui ne se préoccupe pas de fréquence et pour un mot donné ne considère que sa présence - ou son absence - dans le texte considéré. Ou plus exactement, pour deux textes dont on cherche à apprécier la connexion, un mot contribue à rapprocher ces deux textes s'il est commun aux deux et à augmenter la distance s'il est privatif et ne se rencontre que dans un seul. La collection des données est assez lourde parce qu'il faut considérer tous les mots sans exception et que pour chacun on doit prendre en compte tous les appariements de textes deux à deux (le nombre des confrontations pour n textes étant égal à n * (n-1) / 2). Pour chaque paire considérée, la distance obtenue tient compte de l'étendue de l'un et l'autre vocabulaires, selon la formule :
d = ((a-ab)/a) + ((b-ab)/b),
où ab désigne la partie commune aux vocabulaires a et b (a-ab et b-ab recouvrant les parties privatives). Chacun des deux quotients (dont la somme constitue la mesure de la distance) est le rapport, pour un texte donné, du vocabulaire exclusif au vocabulaire total. Il évolue nécessairement entre 0 et 1. La somme a donc pour limites 0 et 2. En réalité la somme se situe autour de 1 et reste insensible aux différences d'étendue des deux textes mis en parallèle. Le calcul est en réalité un peu plus complexe dans la dernière version d'Hyperbase. Il intègre non seulement les mots communs aux textes A et B et les mots privatifs qui se trouvent dans A sans être dans B et réciproquement, mais aussi les mots du corpus qui ne se trouvent ni dans A, ni dans B. Ces mots pareillement rejetés par les deux textes contribuent dans une certaine mesure à rapprocher, même négativement, les deux textes, puisqu'ils partagent les mêmes répulsions ou les mêmes désintérêts. Ces deux variantes du calcul n'épuisent pas les possibilités de la méthode Jaccard. On a compté jusqu'à vingt autres indices, tous fondés sur les mêmes ingrédients[1].

1. Convergence de deux méthodes d'analyse : factorielle et arborée

Une fois obtenu le tableau des distances, son exploitation peut emprunter deux voies différentes mais convergentes : l'analyse de correspondance (figure 1) et l'analyse arborée (figure 2) :

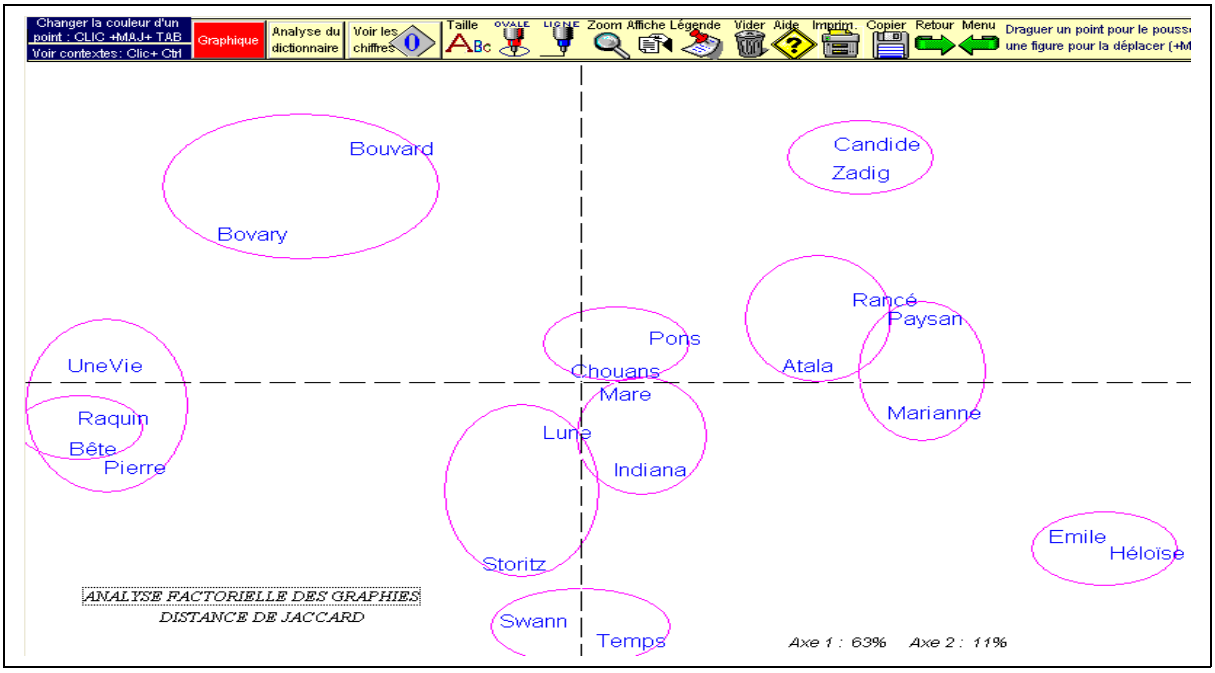
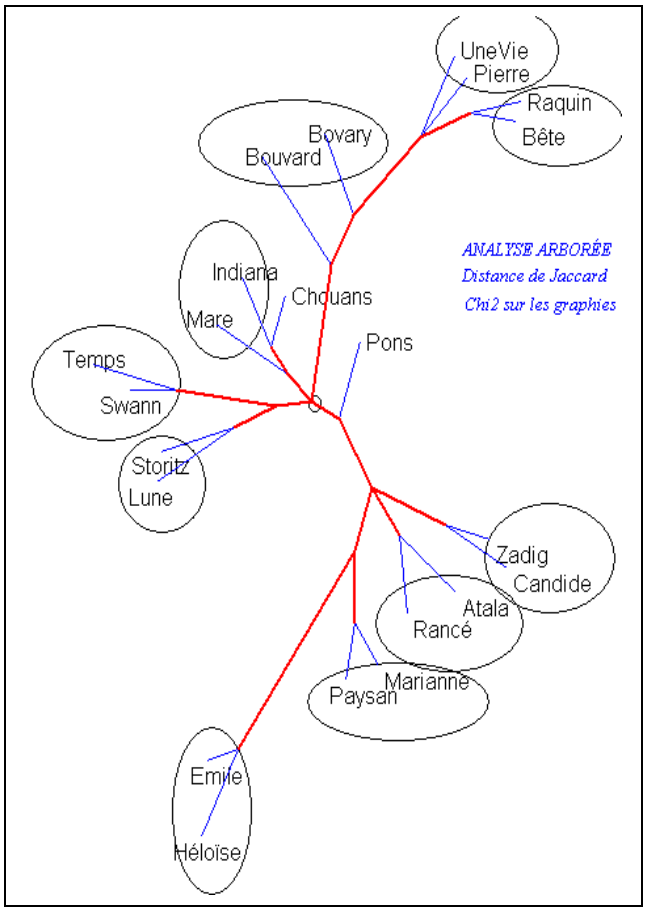


Figure 1. Analyse factorielle de la distance intertextuelle (analyse des graphies, méthode Jaccard)

[1] J.B. Baulieu, « A classification of Présence/Absence Based Dissimilarity Coefficients », *Journal of Classification*, 6:233-246 (1989).

Qu'on ne cherche pas l'influence du genre qui est ici neutralisée : les 22 textes réunis dans ce corpus relèvent tous du genre narratif[2]. Restent deux variables qu'on a voulu croiser : la chronologie et l'écriture propre à chaque écrivain. L'analyse factorielle représentée dans la figure 1 paraît suivre la chronologie, puisque tous les textes du XVIIIe siècle et de la première moitié du XIXe se portent à droite, quand les textes les plus récents campent dans la partie gauche. Il y a pourtant l'exception remarquable de Proust et de Verne qui fuient la compagnie de Zola et des romanciers réalistes et se tiennent à cheval sur l'axe central. On voit ainsi que le tempérament d'un écrivain peut résister à la pression du temps. Pour mettre en évidence la force du lien qui lie chaque texte à son auteur, le corpus incorpore deux textes de chaque écrivain, situés au début et à la fin de sa carrière. Or l'analyse factorielle reconnaît aisément ce lien et place toujours à proximité les textes qui appartiennent à la même plume, ce qu'on peut vérifier dans la figure 1, où les binômes sont clairement identifiables : à droite Marivaux, Rousseau, Voltaire et Chateaubriand, à gauche Flaubert, Maupassant et Zola et au centre Balzac et Sand, et plus bas Verne et Proust.
Or l'analyse arborée, représentée dans la figure 2, donne, plus clairement encore, les mêmes enseignements. Aux deux bouts de la chaîne on rencontre les mêmes configurations que précédemment, avec une zone de transition indécise où flottent les textes de Balzac et de Sand.



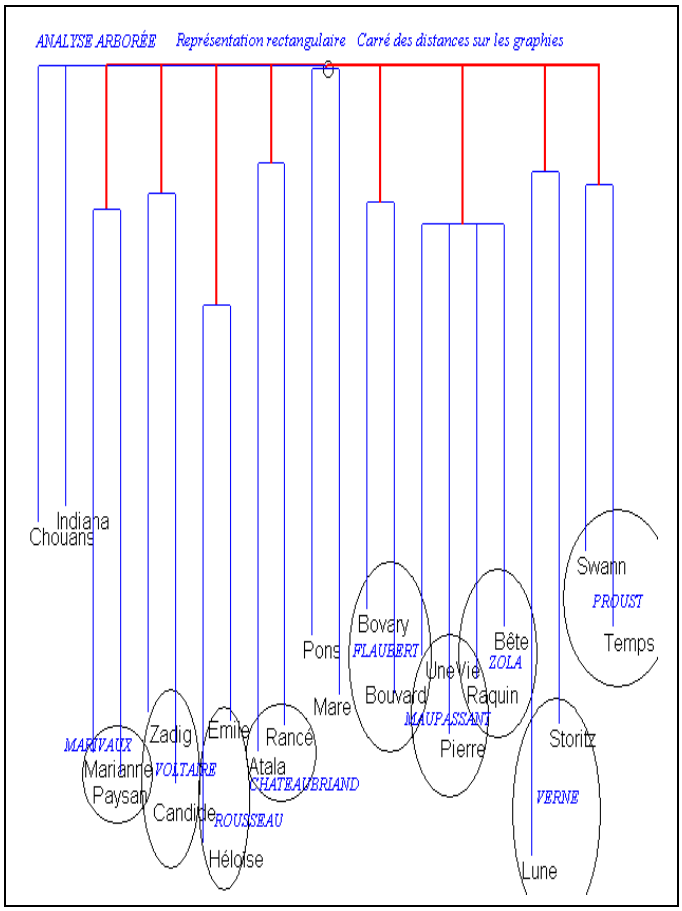


Figure 2. Analyse arborée (graphies, méthode Jaccard) Représentation radiale sur Chi2 des distances

Figure 3. Même analyse (graphies, Jaccard) Représentation rectangulaire sur le carré des distances

Et de la même façon Proust et J. Verne se tiennent à l'écart sur une branche latérale. Ici toute l'information du tableau des distances se trouve résumée au mieux, alors que dans l'analyse factorielle les deux premiers facteurs n'épuisaient pas l'inertie.

[2] Composition du corpus : Marivaux : *La Vie de Marianne* et *Le Paysan parvenu*, Rousseau : *La Nouvelle Héloïse* et *Émile*, Voltaire : *Zadig* et *Candide*, Chateaubriand : *Atala* et *La vie de Rancé*, Balzac : *Les Chouans* et *Le cousin Pons*, Sand : *Indiana* et *La mare au Diable*, Flaubert : *Madame Bovary* et *Bouvard et Pécuchet*, Maupassant : *Une Vie* et *Pierre et Jean*, Zola : *Thérèse Raquin* et *La Bête humaine*, Verne : *De la terre à la lune* et *Les secrets de Wilhelm Storitz,* Proust : *Du côté de chez Swann* et *Le Temps retrouvé*.

2. Convergence données brutes / données pondérées

La figure 3 offre une variante de l'analyse arborée. Il n'y aurait pas lieu de s'étonner qu'il y ait recoupement des deux représentations, radiale et rectangulaire, si les données étaient exactement les mêmes. Le tableau des distances absolues est bien commun aux deux analyses mais dans un cas (figure 2) il a été pondéré et transformé en profil, grâce au calcul du Chi2, dans l'autre (figure 3) il a été accentué, chaque distance étant portée au carré. Or, pondérées, amplifiées, ou brutes, les données s'organisent de la même façon. L'avantage de la transformation est toutefois de rendre plus claire la décantation et de mieux marquer oppositions et rapprochements.

Mais la méthode Jaccard fait peut-être la part belle aux raretés du vocabulaire et particulièrement aux hapax, au détriment des fréquences plus courantes. Les classes de fréquence élevée perdent ainsi tout poids dans le calcul, puisqu'elles se trouvent nécessairement dans la partie commune et inévitable du vocabulaire (*ab*). Ce calcul peut être jugé trop sensible aux artefacts que peuvent produire l'inconstance de l'orthographe, les fautes de frappe, l'abondance des noms propres, bref tous les phénomènes, parfois mineurs et négligeables, qui engendrent la multiplication des formes. Certains considèrent que c'est donner trop d'importance à l'excentricité et qu'une véritable appréciation de la distance entre deux textes doit considérer, pour un même mot, le dosage des fréquences dans les deux textes comparés.

3. Convergence de deux calculs de distance : Jaccard/Labbé

Or Dominique Labbé, a proposé un algorithme efficace qui pour chaque mot apprécie la distribution réelle des fréquences dans les deux textes A et B en comparant les fréquences observées non plus à la répartition théorique mais à l'écart maximal possible dans cette distribution: $D_{(A,B)} = \sum d_i \ / \sum dmax_i$ pour i variant du premier au dernier mot du vocabulaire.

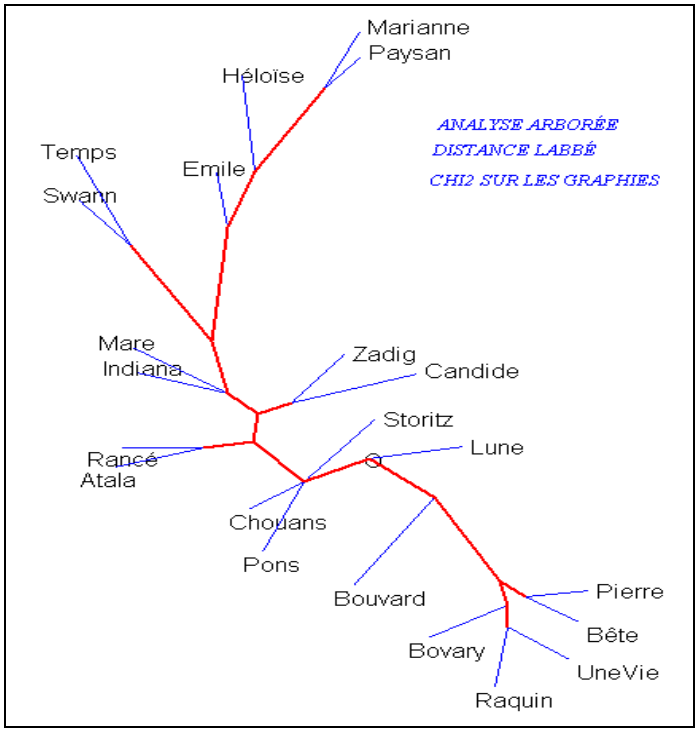


Figure 4. Analyses arborée des graphies. Méthode Labbé

On prendra la mesure de l'écart en rapprochant la figure 4 de la figure 2. Il faut reconnaître qu'en l'occurrence l'écart est faible et que les mêmes lignes de force s'y dessinent, si l'on néglige une inversion verticale qui ne tire pas à conséquence. Même s'il tient compte des moyennes et hautes fréquences, le coefficient de Labbé est surtout sensible, comme celui de Jaccard, aux mots de basse fréquence, qui sont les plus nombreux. Et l'expérience montre que, quel que soit le corpus étudié, les deux coefficients mettent en relief les mêmes influences s'exerçant dans le même sens et avec la même intensité.

Jusqu'ici la convergence des résultats n'est pas en soi un progrès car c'est toujours le même objet qu'on a soumis aux expériences méthodologiques. On a supposé acquis le tableau des distances.

On a seulement varié les éclairages, les prises de vue et le traitement de l'image en laboratoire. Mais le tableau des données peut et doit être remis en question. On n'a considéré que les graphies pour apprécier la distance intertextuelle. Mais les graphies sont un matériau dégradé et désintégré qui ne donnent qu'une image plate et déconstruite du texte. Ne pourrait-on pas affiner le produit en séparant ce qui doit l'être, les homographes, et en rassemblant ce qui doit l'être, les formes qui se rattachent à la même entrée du dictionnaire.

4. Convergence des graphies et des lemmes

On a donc lemmatisé les deux millions de mots du corpus, en utilisant *Cordial*. Le résultat, inscrit dans la figure 5, reflète en miroir l'image de la figure 4. Le recouvrement des deux graphiques est presque parfait. Non seulement les textes d'un même auteur sont placés pareillement à proximité l'un de l'autre, mais les rares décrochages observés dans le graphique des graphies se retrouvent dans celui des lemmes : les deux textes de Flaubert, de Verne et de Rousseau restent proches mais n'ont pas un lien direct. Et dans les deux analyses les textes de Zola et de Maupassant se mêlent les uns aux autres.

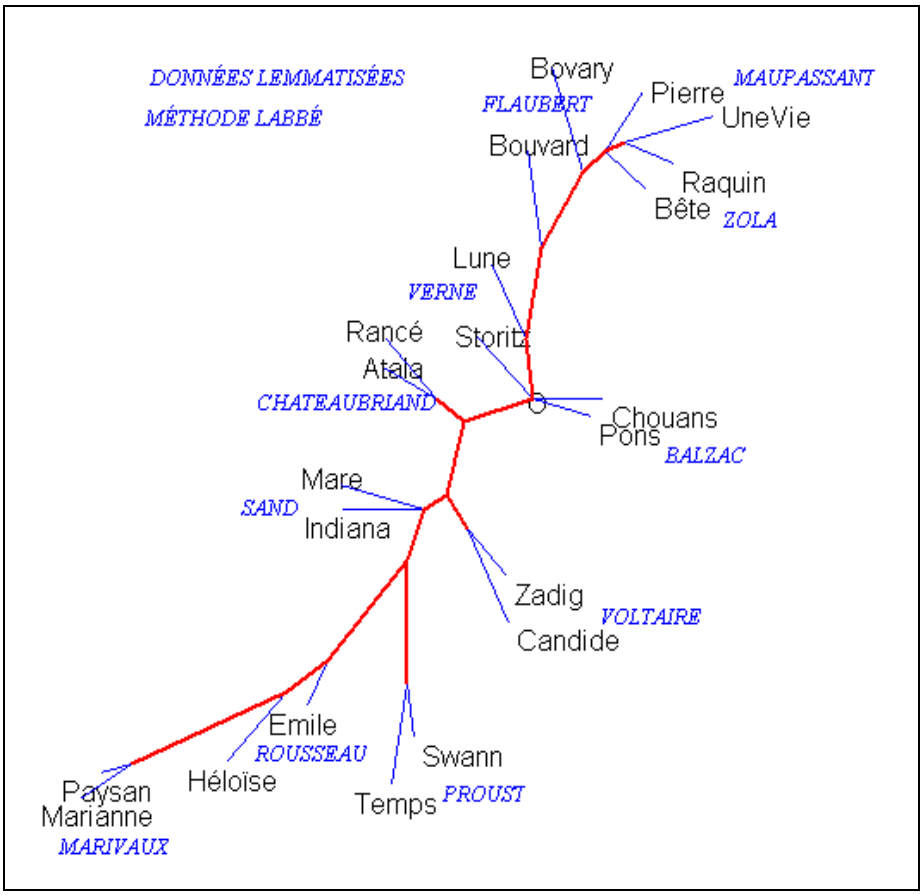


Figure 5. La distance intertextuelle fondée sur les lemmes (méthode Labbé)

5. Convergence des mots et des codes grammaticaux

Ainsi les 22 textes de notre corpus se répartissent de la même façon lorsqu'est mesurée la distance entre leurs vocabulaires, lemmatisés ou non. La distance intertextuelle peut aussi être appréciée en dehors de toute influence thématique, en observant uniquement la distribution des codes grammaticaux ou des structures syntaxiques, indépendamment des mots auxquels ces codes ou structures sont attachés. Comme chacun des quatre niveaux d'observation peut donner lieu à un calcul fondé sur la fréquence (méthode Labbé) ou sur la présence/absence (méthode Jaccard), on dispose en fin de compte de huit points de vue qui heureusement convergent. On préfèrera toutefois la méthode Labbé s'il s'agit des codes car la variété y est limitée et les effectifs importants (figure 6), et la méthode Jaccard pour la raison inverse s'il s'agit de structures (figure 7).

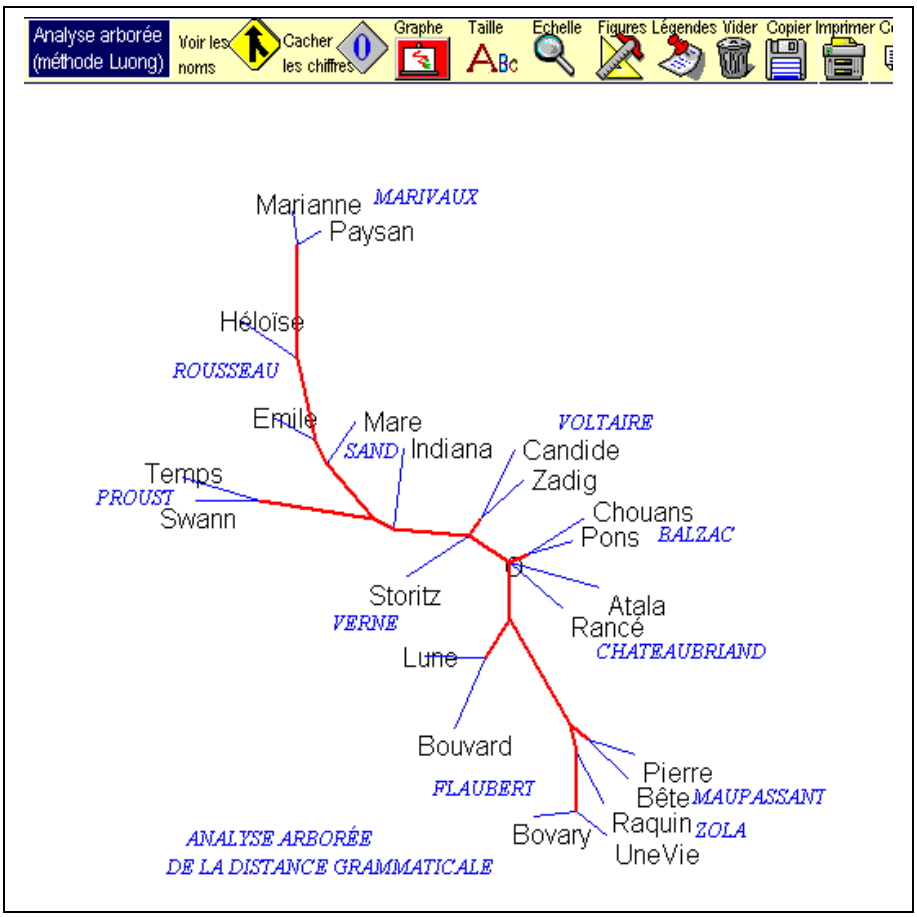


Figure 6. Analyse factorielle des codes grammaticaux (méthode Labbé)

Cette fois la convergence ne laisse pas d'étonner. Si l'on peut comprendre le parallélisme des mots-graphies et des mots-lemmes, car les premiers sont inclus dans les seconds, on ne voit pas *a priori* quel lien nécessaire pourrait être établi entre les lemmes et les codes grammaticaux. Dans l'effectif du lemme *aimer*, il y a certes l'imparfait *aimait*, mais aussi toutes les autres formes du verbe. Et dans l'effectif du code *imparfait 3e personne du singulier* il y certes la forme *aimait*, mais aussi des centaines d'autres, comme *avait*, *était*, parfaitement étrangères au verbe *aimer*. Lemmes et codes grammaticaux se présentent en principe comme des variables indépendantes, les premiers plutôt thématiques et les seconds plutôt stylistiques. Les uns et les autres sont pourtant traversés par des courants semblables, où le tempérament des écrivains se manifeste, tantôt cédant, tantôt résistant à la dérive du temps. L'acte d'écrire ne s'exerce pas en deux temps, le choix du style succédant à celui du thème, comme on fait pour l'achat d'une voiture, la sélection de la couleur venant après celle du modèle. L'écriture implique un choix simultané, cohérent quoique souvent inconscient, des variables thématiques et stylistiques.

6. Les structures syntaxiques

La désincarnation est poussée plus loin encore dans la figure 7 qui est relative aux structures syntaxiques et où la charge sémantique des mots et des textes est complètement évacuée. Tous les schémas syntaxiques rencontrés entre deux ponctuations sont relevés dans le corpus, catalogués et cumulés. Ce qu'on prend en compte n'est plus le dosage des parties du discours, mais leur assemblage et le rythme de la segmentation. Quelques particularités apparaissent comme le déplacement de Proust et Voltaire. La longue phrase du premier le rapproche de l'époque classique, alors que le second échappe à son temps et annonce la modernité. Mais ces retouches de détail ne perturbent guère le mouvement d'ensemble qui reproduit les figures précédentes.

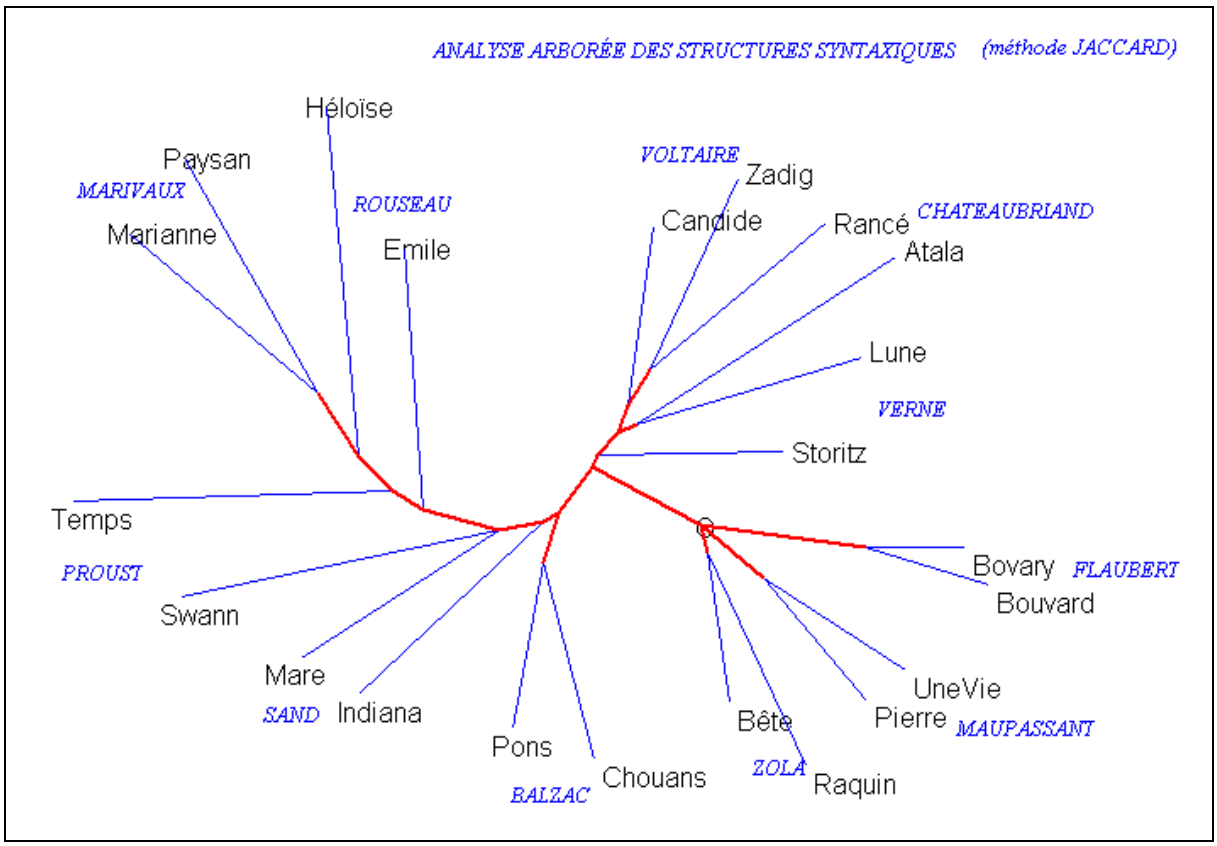


Figure 7. Analyse factorielle des structures syntaxiques (méthode Labbé)

7. Les codes sémantiques

Selon qu'on utilise les filtres appropriés, à l'échelle macroscopique ou élémentaire, un corpus peut être envisagé comme un ensemble de textes, ou de phrases, ou de graphies, ou de lemmes, ou de codes grammaticaux, ou de structures syntaxiques, ou de balises sémantiques. Ce dernier niveau est le plus difficile à atteindre, car le sens des mots échappe en grande partie à la machine. Le lemmatiseur *Cordial* propose pourtant une ontologie extérieure qui distribue des étiquettes sémantiques aux mots-pleins du corpus. Cet étiquetage est discutable, certaines balises sont curieusement nommées et leur attribution est approximative. Pourtant l'analyse factorielle appliquée à de telles données reprend pour l'essentiel la typologie observée dans le corpus. Comme précédemment, la figure 8 oppose aux autres les représentants du roman réaliste et naturaliste. Elle ajoute toutefois une information précieuse : la carte des thèmes est superposée à celle des textes et l'interprétation s'en trouve facilitée. Car la proximité d'un texte et d'un thème acquiert une signification. Ainsi les œuvres de Flaubert, Maupassant et Zola sont tournées vers la description du milieu et des réalités concrètes, matérielles, corporelles, qu'on devine derrière les thèmes qui les entourent : *concret, quotidien, production, corps, sens, santé, vivant, espace, cinétique* (= mouvement). À l'opposé la littérature classique (c'est là aussi que Proust prend place) se préoccupe davantage des réalités morales, psychologiques, religieuses, sociales ou politiques (*éthique, spiritualité, droit, volonté, esprit, homme, économie, pouvoir, société*).

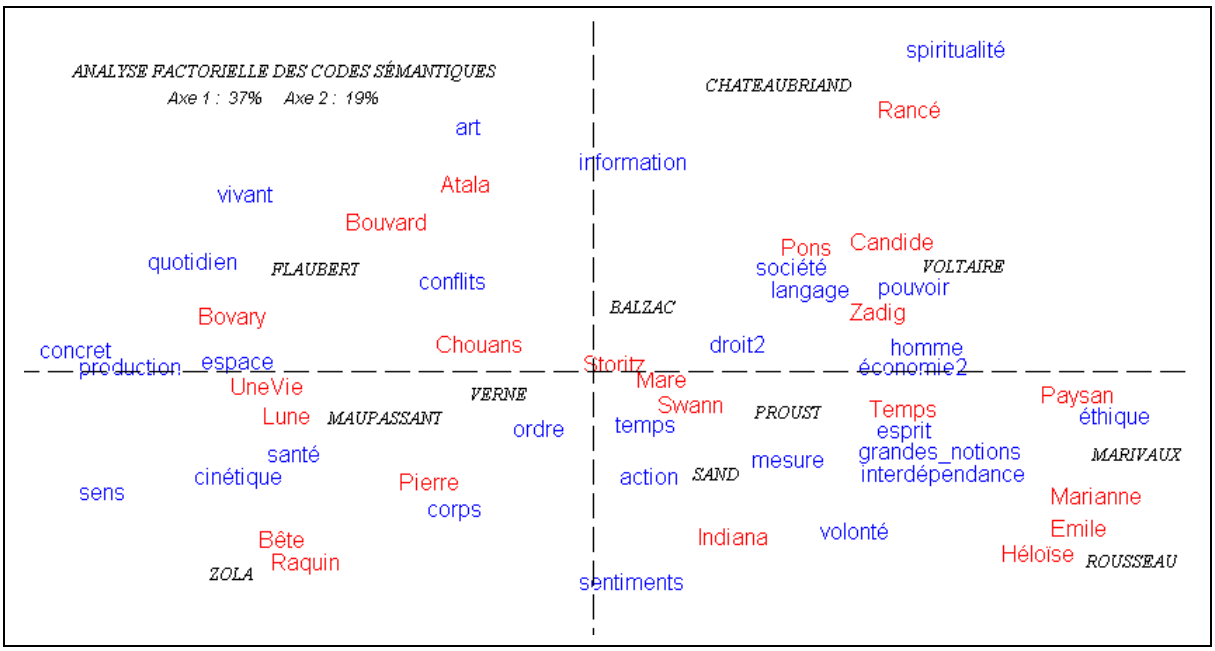


Figure 8. La carte thématique du corpus

8. L'expérience des n-grammes

Chacun des jalons sur lesquels s'appuie la segmentation peut être neutralisé ou déplacé, qu'il s'agisse de ceux qui séparent les textes, ou les phrases, et même des séparateurs qui isolent les mots. Imaginons un Champollion devant une immense pierre de Rosette où les mots n'auraient pas de frontières distinctes et qui contiendrait les dix millions de caractères de notre corpus. Et supposons que devant ce texte inconnu on ait promené une loupe de proche en proche en isolant quatre lettres à la fois, et en déplaçant la fenêtre d'une seule case à chaque pas. Ainsi le mot *fenêtre* génèrera quatre n-grammes successifs : *fenê*, *enêt*, *nêtr* et *être,* qui tiendront lieu de « mots ». Cette fois, au lieu d'enrichir le texte en le dotant de codes grammaticaux ou sémantiques, on l'appauvrit jusqu'à le rendre illisible. Le blanc ayant disparu, les mots ne sont plus reconnaissables, n'ayant ni queue ni tête. Et pourtant ce rébus opaque ne pose aucun problème au programme de reconnaissance, qui retrouve les textes issus de la même plume et dresse une carte d'attribution aussi claire que celle des lemmes. Le graphique obtenu sur ces données perverties est superposable en tous points à ceux que le matériau linguistique épuré avait produits (notamment les figures 2, 4, 5 et 6). Avant de nous interroger sur cette stabilité étonnante des traitements statistiques, poursuivons jusqu'à l'extrême notre jeu de déconstruction. Après avoir cassé les mots, cassons l'alphabet. Négligeons les accents et oublions les lettres. Procédons comme une sténo indolente ou inculte qui n'aurait à sa disposition que deux symboles pour noter ce qu'elle entend : le signe V pour une voyelle et le signe C pour une consonne.

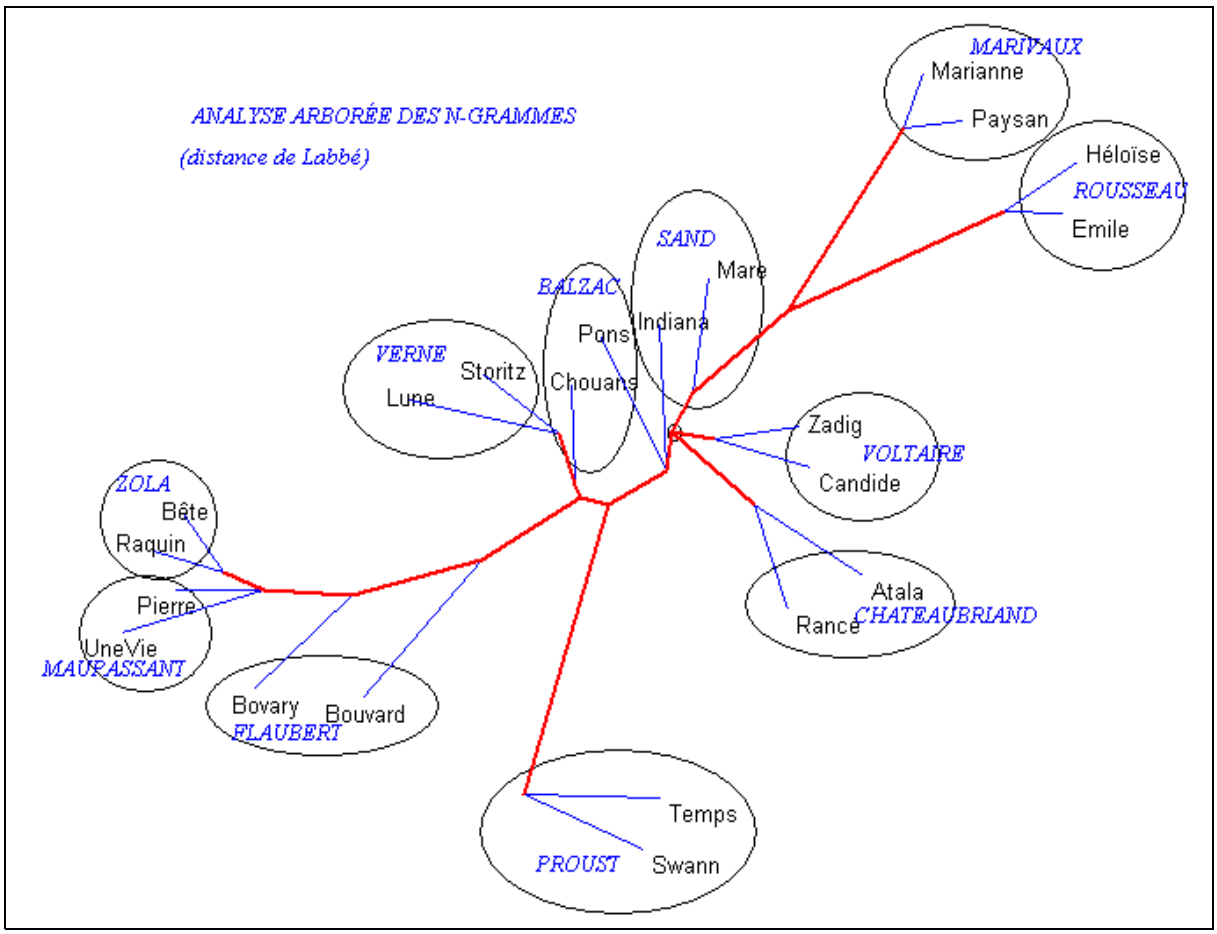


Figure 9. Analyse arborée des n-grammes

9. L'expérience ultime Consonne/Voyelle

Le résultat de cette réduction drastique apparaît ci-dessous : difficile de reconnaître la dernière ligne de la *Recherche du temps perdu* dans cette suite de CV.

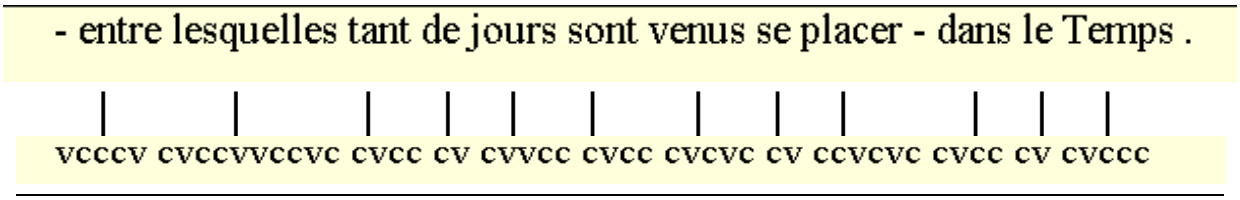


Figure 10. La dernière ligne du Temps retrouvé *réduite à une succession de voyelles et de consonnes*

La perte d'information semble irrémédiable : tous les mots de deux lettres n'ont le choix qu'entre trois combinaisons, CV, VV, et VC. Ceux de trois lettres ont un choix à peine plus ouvert. C'est dire que tous les mots-outils sont quasiment confondus. À elles seules les trois premières combinaisons, à savoir CV, CVC et VC, représentent le tiers de la surface imprimée. Et inversement, avec un alphabet aussi pauvre, les combinaisons rares se raréfient encore. Il n'y a plus que 607 hapax, contre 19156 dans le texte original. Et pourtant le miracle se produit : imperturbable, la machine arrive à démêler le nœud gordien et à proposer une typologie des textes qui s'écarte à peine de celles qu'on a obtenues avec un matériau cent fois plus riche et plus précis. La plupart des binômes ont un lien direct, ce qu'on observe pour Marivaux, Voltaire, Chateaubriand, Balzac, Maupassant et Proust. Ailleurs la liaison est courte même si elle n'est pas immédiate. Le mouvement d'ensemble est grossièrement respecté. Les textes du XVIIIe forment un bloc, ceux du roman réaliste un autre, et l'irrédentisme de Proust, à l'écart sur une branche latérale, est bien visible.

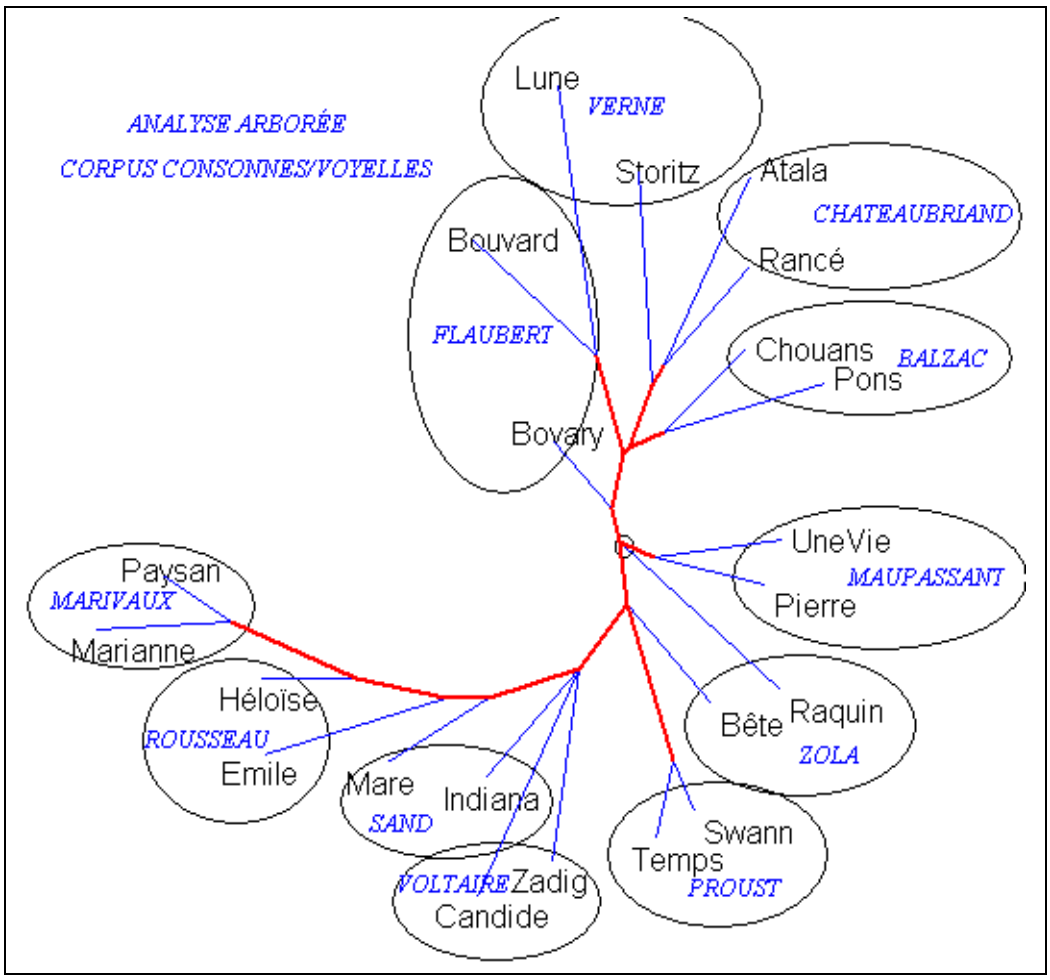


Figure 11. Analyse arborée du corpus consonnes/voyelles

Réduit à une combinaison de trois éléments – consonne, voyelle et blanc – le corpus prend l'allure du génome et les mêmes méthodes de décryptage pourraient s'y appliquer. Jean-Pierre Anfosso[1] dans sa thèse a montré qu'on pouvait y parvenir, quand l'emploi d'un alphabet restreint autorise le traitement des chaînes de Markov.

L'appareillage statistique, appliqué au langage, apparaît ainsi d'une remarquable stabilité, jusqu'à provoquer le soupçon que c'est toujours la même chose qu'on mesure. Il produit des résultats convergents à des niveaux très éloignés et très variés, depuis les regroupements ontologiques les plus larges - et les plus flous - jusqu'aux analyses les plus microscopiques des molécules et des atomes du langage. Les mêmes lignes de force s'y reconnaissent, quelle que soit la focale utilisée ou l'éclairage ou l'angle de la prise de vue. Le corpus est comme une boule : qu'on le considère d'en haut ou d'en bas, de la droite ou de la gauche, de l'avant ou de l'arrière, l'image est la même.

[1] J.P. Anfosso, *Contribution à une modélisation statistique du langage et à sa mise en œuvre informatique*, Nice, nov. 2002.

TYPOLOGIE DES CONCEPTS DE LINGUISTIQUE : ÉVALUATION ET ÉLABORATION EN CORPUS DE CRITÈRES DISCRIMINANTS

Céline POUDAT
CORAL, Université d'Orléans

SOMMAIRE

1. Introduction

Le présent article propose un ensemble de critères typologiques pour discriminer les concepts de linguistique, éprouvé en corpus à partir d'une collection homogène en genre de 224 articles extraits de 32 numéros de revues (soit 11 revues) francophones de sciences du langage essentiellement publiés autour de 2000.

Si les dictionnaires sémantiques, les ontologies et les bases terminologiques sont en pleine expansion, la description des concepts scientifiques en tant qu'unités textuelles et textualisées est encore peu développée ; le contexte des termes ou des concepts est certes pris en compte par les dictionnaires contextuels, et on recense bien des travaux dans lesquels la sélection des termes est objectivée en corpus (par une analyse des spécificités par exemple, L'homme, 2004), mais peu d'entreprises cherchent à caractériser les concepts en tant qu'unités méso-sémantiques (palier intermédiaire entre la micro-sémantique du mot et la macro-sémantique du texte) ; dans cette perspective, qui a montré son intérêt et son efficacité dans plusieurs études récentes (Loiseau 2003, Valette 2003), les concepts, que l'on peut décrire comme des thèmes, sont potentiellement corrélés à des marqueurs ou à des formes expressives relevant de tous les niveaux de l'analyse linguistique (Rastier, 2003).

L'analyse thématique des concepts scientifiques étant encore à son stade exploratoire, on manque encore de critères pour discriminer les objets et les formes sémantiques. La présente étude vise ainsi à éprouver différents critères typologiques en corpus : fréquence et répartition des concepts dans les textes du corpus, corrélations morphosyntaxiques, co-occurrents lexicaux, incidence du style d'auteur et du numéro thématique de revue et configurations tactiques. On insistera sur le caractère non isolé des critères proposés, qui discriminent d'ailleurs souvent les mêmes phénomènes, mais sur des plans distincts.

Bien que le présent article porte sur un corpus génériquement et domanialement homogène, on peut penser que certains des critères discriminants établis sont généralisables à d'autres disciplines des sciences humaines.

2. Fréquence et répartition

Supposés constituer un mode d'accès privilégié aux thèmes scientifiques linguistiques, ce sont les substantifs les plus représentés que nous avons choisi d'extraire. Etant donné le peu de corpus de comparaison disponibles, le recours à une analyse des spécificités a dû être écarté[1].

[1] Mentionnons toutefois que nous avons par ailleurs mis à jour, au sein d'une étude comparative des discours linguistique, philosophique et critique (Loiseau, Poudat et Ablali, 2006) certains items spécifiques à la linguistique, qui renvoient la discipline à ses observables (*verbe, phrase, énoncé, mot*, etc.), en excluant la *langue* ou le *sens,* que s'approprient également la philosophie et la critique ; il nous semblerait peu pertinent d'exclure certains objets parce que d'autres disciplines les empruntent.

Les variations flexionnelles (singulier/pluriel) ont été prises en compte, dans la mesure où le trait « nombre » n'indique pas seulement la pluralité : *langue* et *langues* sont ainsi deux concepts linguistiques distincts.
Bien que le *texte* l'emporte sur la *phrase* en termes de fréquences absolues (1313 *vs*. 1237), on observe qu'il apparaît pourtant dans un nombre plus restreint d'articles (121 *vs*. 143), et que c'est finalement *l'énoncé* qui domine selon ce dernier critère (149 textes) : on voit là l'intérêt de prendre la fréquence *et* le nombre de textes d'apparition de l'occurrence.
Nous avons donc ordonné l'ensemble des substantifs en prenant en compte les deux paramètres, ce qui nous donne le classement suivant (seuls les 20 premiers noms communs au singulier sont présentés dans le tableau qui suit) :

Rang	Substantif	Fréquence absolue	Textes
1	sens	2136	200
2	forme	1840	208
3	cas	1693	214
4	langue	2037	176
5	type	1650	205
6	relation	1654	183
7	objet	1500	191
8	point	1297	210
9	discours	1687	155
10	contexte	1568	157
11	rapport	1223	196
12	analyse	1263	185
13	verbe	1632	142
14	sujet	1351	165
15	fait	1057	194
16	fonction	1031	187
17	question	986	191
18	énoncé	1206	149
19	phrase	1237	143
20	partie	917	190

Graphique : Substantifs au singulier ordonnés par fréquence et par textes

On distingue globalement deux types de substantifs : les candidats concepts de linguistique (en gris), et les substantifs relevant visiblement de la méthodologie scientifique à l'œuvre dans les articles. On observe ainsi un intérêt prononcé de la discipline pour le *sens*, qui détrône la *langue*, objet pourtant intuitivement premier de la linguistique. Le *discours* est également très honorablement représenté, tandis qu'on observe un intérêt particulier pour le *verbe*, dont la forme fléchie plurielle détient le second rang des substantifs au pluriel (1225 occ. pour 117 textes).
Les substantifs restants relèvent de la logique (*relation, rapport*), de la typologie (*cas, type*) ou sont trop ambigus pour renvoyer directement à un objet linguistique (*sujet, objet, fonction, point*).
Les substantifs relevés au pluriel corroborent généralement le classement précédent, bien qu'on observe certains substantifs résolument déterminés en nombre : ainsi, « sens » est le substantif singulier le plus relevé dans le corpus (2186 occ. au singulier) et il est notable qu'il soit globalement peu employé au pluriel (179 occ. / 83^{e} rang). Il en va de même pour *discours* (9ème substantif), qui est relevé 1732 fois au singulier et seulement 146 fois au pluriel (118^{e} rang), ou encore pour *langage* (1208 occ. au singulier *vs*. 30 au singulier). S'il est intuitif que *le* discours et *les* discours, ou *le* langage et *les* langages ne renvoient pas aux mêmes concepts linguistiques, de tels écarts demeurent surprenants.
On ne note pas de différences aussi importantes à l'inverse, mais plusieurs substantifs qui relèvent davantage de la méthodologie que d'une thématique strictement linguistique, sont essentiellement pluriels : *données, conditions, caractéristiques, phénomènes, traits, critères, résultats* ou *contraintes*.

Soulignons que ce bref panorama des candidats concepts linguistiques ne porte que sur les *hautes fréquences*. Malgré leur intérêt descriptif, les hapax et les éléments moins – ou plus inégalement – représentés ont été globalement écartés des analyses qui suivent, dans la mesure où ils se prêtent difficilement à l'analyse statistique.

3. Co-occurrents et corrélats lexicaux : exploration de trois paliers de régulation linguistique

L'examen des co-occurrents lexicaux est particulièrement crucial dans le processus de qualification thématique des candidats concepts : si leur nombre implique une plus ou moins grande stabilisation de la forme et si les écarts observés mesurent le degré de corrélation contextuelle entre les mots, leurs éventuelles intersections sémiques – qui indique la présence d'isotopies – les font accéder au statut de *corrélats sémantiques* (Rastier).

Cette méthode, qui a montré sa pertinence dans l'examen des textes littéraires, linguistiques et philosophiques (*e.g.* Valette 2003, Bourion 2001, Loiseau 2003), permet de faire émerger des phénomènes linguistiques non-, voire contre-intuitifs, qui sauraient difficilement être appréhendés par d'autres biais. On soulignera qu'en matière de textes linguistiques ou philosophiques, elle a surtout été éprouvée au palier individuel de l'auteur (Deleuze chez Loiseau et Guillaume chez Valette).

Bien que le niveau de normalisation du *style d'auteur* soit naturellement pertinent pour évaluer l'appropriation singulière des concepts d'un domaine ou d'une discipline scientifique, deux paliers de niveaux supérieurs nous semblent également significatifs : celui du *genre*, et nous limiterons nos investigations à l'*article de revue*, et celui du *thème*, ou du *numéro thématique de revue* si l'on s'intéresse au domaine linguistique – la plupart des numéros de revue étant organisés autour d'une thématique ou d'une problématique fédératrice.

Nous adopterons ainsi une démarche progressive et descendante (de la généralité du genre à la singularité du style), en parcourant ces trois paliers de normalisation du discours scientifique linguistique. Dans la mesure où il n'est pas envisageable ici d'analyser les co-occurrents de l'ensemble des candidats concepts du corpus, nous avons choisi de nous concentrer sur les objets *sens* et *langue,* choix motivé par leurs hautes fréquences et leur qualité d'objets / objectifs de descriptions linguistiques privilégiés.

3.1. Le palier générique (corpus ASLF)

On notera d'abord qu'en dépit de leurs fréquences distinctes (2471 occ. de *sens vs*. 1798 occ. de *langue*), *sens* et *langue* ont quasi le même nombre de co-occurrents : 209 pour *sens vs*. 207 pour *langue*. On relève des écarts de corrélation plus importants pour *langue* que pour *sens* : *langue* draine ainsi plus de lieutenants stabilisés que *sens* (*e.g. langue des signes, parlée, française, maternelle, naturelle, étrangère, usuelle, de spécialité, courante*, etc.), eux-mêmes corrélés à des co-occurrents renvoyant à des contextes spécifiques.

Ce phénomène entraîne des difficultés d'évaluation dans la mesure où les premières corrélations obtenues ne sont pas nécessairement caractéristiques du corpus : elles peuvent facilement être liées à l'un de ses sous-ensembles singuliers (un numéro thématique le plus souvent). Par exemple, *langue* apparaissait d'abord corrélée à *signes, sourds, LSF, entendants*, etc., éléments bien spécifiques à *un* numéro thématique singulier du corpus, dédié à la *langue des signes*.

On voit là toute la difficulté que pose l'observation des formes de haute fréquence, qui drainent de nombreux figements qui s'autonomisent de la notion première.

On observe le même phénomène pour *sens*, qui n'a qu'un rôle de formant dans de nombreux figements : *sens littéral / sens figuré, sens strict / sens large, sens commun…* auxquels on peut adjoindre *immanence du sens, sens distributionnel, donation* du sens*,* [unité] *porteuse* de sens, *compositionnalité du sens…* si l'on restreint le contexte pris en compte du paragraphe aux 50 caractères avoisinant la notion.

Notons que les deux objets sont corrélés à *Saussure*, et que la *langue* est corrélée à la dichotomie saussurienne *langue / parole* (*parole* et *Saussure*) tandis que *sens* s'inscrit dans la problématique du *signe* (*signifié, signe, signifiant, Saussure*).

3.2. Le niveau du numéro thématique

Comme nous avons déjà pu l'observer avec *langue / langue des signes*, le numéro thématique de revue participe substantiellement à la régulation thématique du genre et du domaine linguistique.

Le différentiel entre le niveau du genre et celui du numéro thématique permet ainsi de mettre à jour les dominantes thématiques spécifiques des numéros et à terme, d'évaluer ce qui constituerait un noyau dur conceptuel de la discipline linguistique.
Si l'on prend par exemple un numéro thématique spécifique, la revue *Contexte(s)*[1] ici, on observe que la présence du lexème *contexte* dans les premiers co-occurrents de *sens* s'avère en grande partie liée au numéro *Contexte(s)* : au sein du corpus *Contexte(s)*, l'item *contexte* est le cinquième co-occurrent de *sens*, tandis qu'il n'est que le 63e lorsqu'on considère le reste du corpus d'articles.

Premiers co-occurrents de *sens* : Sens (34.66), classique (6.06), établissement (5.60), le (5.50), contenu (5.50), littéral (5.35), contexte (5.33), au (5.19), virtuel (4.80), commun (4.77), hors (4.77), notion (4.64), message (4.51), mot (4.46), approche (4.43)

Les co-occurrents de *sens* diffèrent d'ailleurs fortement aux niveaux de la revue et du reste du corpus : on ne retrouve pas la plupart des items obtenus d'un corpus à l'autre. Le concept de *sens* dans la revue observée ne semble pas débattu en tant que tel, mais relativement à la notion de contexte ; en d'autres termes, on ne s'intéresse au *sens* que par rapport au *contexte*, d'où le figement *sens littéral* (en contexte zéro) et la récurrence de la notion de *contenu* (topos : le sens n'a pas de contenu sans prise en compte du contexte) :

> Lakoff et Johnson (La métaphore dans la vie quotidienne 1981, p. 21) montrent, à travers la métaphore du conduit, que l'idée d'un **sens contenu** dans une forme linguistique est inhérente à notre perception de la langue : on parle de « faire passer, donner une idée, introduire une idée dans une phrase, d'une phrase vide de **sens**. » (Les contextes de contexte La notion de contexte dans les Page: 24 c (41ème occ.))
>
> […] il importe de dégager les processus qui permettent l'établissement d'un **sens** en soi, **contenu** dans la langue, indépendamment du contexte. (Production et interprétation du sens : la notion de contexte Page: 279 c (179ème occ.))

On observe également une co-occurrence de *sens* et de la préposition *hors*, qui renvoie encore une fois au contexte, sous des formes différentes (*hors contexte, hors circonstance, hors langue, etc.*).
Contrairement à *sens*, qui semble être un concept finalement caméléon, dont les co-occurrents – et les acceptions – varient selon l'objet observé, *langue* est un concept relativement plus stabilisé, qui semble invariablement porter les traits [Saussure], [système] [française] et [code]. On observera toutefois que l'objet *contexte* ne semble pas spécifiquement recourir à la dichotomie *langue / parole*, la *parole* étant absente des co-occurrents de *langue* dans l'ensemble du numéro.

3.3. Le palier du style

Afin de parfaire et d'approfondir notre description de la stabilisation des deux concepts et de leurs co-occurrents par paliers, nous avons cherché à observer leur stabilité d'un auteur à l'autre, à partir d'un corpus de 118 articles de 12 linguistes français accrédités dans le champ[2].
Les deux concepts sont d'abord très inégalement employés d'un auteur à l'autre : Kleiber recourt à *sens* 13 fois plus que Combettes, tandis que Bergounioux emploie *langue* 25 fois plus que Rabatel.
Le concept de *langue* semble encore une fois plus stabilisé que celui de *sens*, dans la mesure où ses co-occurrents varient peu d'un corpus à l'autre – les deux corpus d'étude étant pourtant bien distincts : la *langue* est ainsi invariablement associée à *linguistique, Saussure* et *système* et aux figements *langue française, langue maternelle* et *langue parlée*, qui apparaissent d'ailleurs plus corrélés à certains auteurs (respectivement Bergounioux, Neveu et Authier).
Outre ces figements, on observe des dominantes thématiques chez certains auteurs, qui renvoient à leurs intérêts et cadres de recherche généraux : *langue* s'inscrit dans une perspective historique chez Bergounioux (*histoire, était, diachronique, etc.*), diachronique chez Combettes (*état(s), ancienne, anciens, catégories, textes, système, trace, surviennent, permettent, moderne,*

[1] Schmoll (*ed.*), Scolia vol. 6, Strasbourg, 1996.
[2] Authier, Barbéris, Bergounioux, Combettes, François, Kerbrat, Kleiber, Moirand, Neveu, Rabatel, Rastier et Siblot. Soulignons que nous avons mis à jour dans des études précédentes (Poudat et Rinck 2005, 2006) le caractère significatif des styles en linguistique aux niveaux morphosyntaxique et lexical.

existence, faits, phénomènes, corpus, vestiges, etc.) ou didactique chez Rabatel (*élèves, maîtres, initiation, grammaire, appris, école, aider, etc.*).
Certains cadres semblent d'ailleurs plus définis et plus rigides que d'autres : il en va ainsi du cadre praxématique dans lequel s'inscrivent Barbéris et Siblot : chez Barbéris par exemple, *sens* est le formant du concept *production de sens.* Le *sens* s'inscrit ici dans une conception praxématique et il articule les objets linguistiques (*préposition, complément, verbes, lexical*) au praxématique (*production, programmes, état, producteur, gestalt, praxémique etc.*). La *langue* (*parlée*) s'inscrit également dans un cadre praxématique, d'où les co-occurrents linguistiques et marxistes *classes, discours, légitime, dominante, souterraine, registres, populaire, etc.* Bien que la *langue* soit discutée en termes sémiotiques saussuriens (*signes, Saussure, nomenclature, système, etc.*), et aux côtés du *langage* (*langage, langagières*) chez Siblot, on observe une conception praxématique du *sens* (co-occurrents *production, praxème, récepteur, procédures, praxis, sociales, programmes, produire*) appliquée cette fois à l'objet *texte* (*texte(s), clôture*).
Le concept de *sens* varie ainsi de manière plus significative et apparaît comme plus discuté que celui de *langue*, qui est généralement corrélé à l'ensemble des items tissant le système conceptuel général de l'auteur : on le trouve en effet associé à des éléments plus spécifiques participant à sa définition ou à sa discussion. Ainsi, il s'inscrit dans la problématique du *signe* chez Bergounioux, qui le distingue du *signifié* (8.98, premier co-occurrent observé) ; on observe ainsi un réseau conceptuel de corrélats articulés autour de cette question : *union* (signifiant/signifié), *unité*, *signe*, *substitution* (de sens à signifié), *signifiant*, *Saussure*, etc. Outre cette isotopie, *sens* est significativement corrélé à *homme* (6.52, rang 2). *Sens* semble ainsi permettre d'articuler l'homme au signe, et par extension, à la langue et au langage.
Si *sens* et *signifié* s'interdéfinissent chez Bergounioux, on le trouve aux côtés de *signification* chez Rastier (premier co-occurrent relevé, écart de 15.59) – *sens* et *signification* partagent d'ailleurs les mêmes corrélats, ce qui indique bien un emploi concomitant des deux notions. *Sens* s'inscrit ainsi dans un environnement sémiotique (*contenu, carré (sémiotique), signifié, signe*) et sémantique – on relève ainsi des éléments qui renvoient à différents modèles du sens (*contenu, mimesis, immanent, texte*…) ; bien qu'originellement distincts et susceptibles de polémiques, ces éléments sont conciliés dans le cadre sémantique rastiérien.
Enfin, on trouve *sens* associé à un sème [+ instable] chez Authier, lié à la perspective énonciative de l'auteur, d'où *risque, (référence) actuelle, ponctuellement, fixité (du signal), équivoque, paraphrasable, polysémie, effets (de sens),* etc., tandis qu'il apparaît particulièrement stabilisé chez Kleiber, et quasi-exclusivement associé à des éléments partageant le sème [+stable] ou [+déterminé] : (*sens) descriptif* (8.50) – qui est d'ailleurs quasi figé chez l'auteur, *détermine* (7.19), *conventionnel* (6.57), *dénominatif* (6.49), *stable* (6.16), *représentationnel* (6.02), *déterminé* (6.02), *conditions* (5.96), *psychologique* (5.96), *codé* (5.62), *stables* (5.36), *instructionnel* (5.36).
On observe ainsi que *sens* constitue un meilleur point d'entrée dans les systèmes conceptuels développés par les auteurs observés : cette distinction pourrait éventuellement participer à celle de concepts de fond / concepts de forme proposée par (Rastier, 2003). *Langue* serait ainsi un concept de fond disciplinaire peu débattu tandis que *sens*, qui paraît moins stabilisé, serait une forme plus discutée.
De manière générale, on peut opposer les concepts dont les co-occurrents sont d'autres concepts des entrées dont les co-occurrents sont des exemples ou des objets de description linguistique : dans le premier cas, le concept semble (inter)défini, voire peut-être débattu, tandis que dans le second, il semble essentiellement instrumental ou méthodologique.
Si on ne considère que les concepts interdéfinis, on peut *a fortiori* opposer les concepts inscrits dans des cadres théoriques et méthodologiques (*e.g.* Praxématique, SI, Sémantique de la référence klébérienne, et cadre énonciatif d'Authier) des concepts ponctuellement débattus où le concept reste de faible fréquence, ou n'est pas pivot dans un système conceptuel – *e.g.* cadre historique de Bergounioux (sens/signe).

4. Corrélations morphosyntaxiques

Nous avons ensuite retenu une sélection de substantifs parmi les hautes fréquences relevées, auxquelles ont été par hypothèse adjoints *sémantique*, *énonciation* et *cotexte* :

SENS : SENS :Nsg, SENS :Npl
SEMANTIQUE : SEMA :Nsg, Npl, SEMA :Asg, Apl

LANGUE : LGUE :Nsg, Npl
DISCOURS : DISC :Nsg, Npl, Asg, Apl (discursif (ve) (s))
PAROLE : PAR :Nsg
LANGAGE : LANG Nsg, Npl, Asg, Apl (langagier, langagière (s))
TEXTE : TXT : Nsg, Npl, Asg, Apl (textuel(le)(s))
CORPUS : CORP Nsg, Npl
INTERPRETATION : INTER Nsg, Npl, Asg, Apl (interprétatif, interprétative)
CONTEXTE : CONT Nsg, Npl, Asg, Apl (contextuel(le)(s))
COTEXTE : COT Nsg, Npl
ENONCIATION : ENONC Nsg, Npl, Asg, Apl (énonciatif(ve)(s))

Dans la mesure où elles sont fondées sur le même morphème, les formes adjectivales des entrées ont également été prises en compte.
Les corrélations des entrées entre elles et avec le jeu de descripteurs morphosyntaxiques ont été observées et c'est sur les textes entiers (avec exemples et citations) qu'ont été effectuées les analyses. L'entreprise peut paraître discutable, mais il s'avère que si l'on extrait les exemples du corpus, il n'est plus possible de voir si les formes sont employées dans des textes contenant des exemples, ceux-ci étant particulièrement identifiables (*v.* Poudat, 2006).
Etant donné les différences d'échelle des deux types de données[1], les substantifs sont globalement corrélés entre eux et peu de corrélations négatives significatives (seuil > -0.2) ont pu être relevées. À l'inverse, plusieurs corrélations positives très élevées (> +0.7) ont pu être observées.
De manière générale, les corrélats des candidats observés diffèrent au singulier et au pluriel, ce qui suppose qu'ils renvoient à des objets distincts. La différence est particulièrement visible si l'on s'intéresse à *sens* : si les deux formes sont corrélées entre elles (+0.33), leurs corrélats diffèrent d'abord en nombre, eu égard à la représentation dix fois plus élevée de *sens* au singulier (14 corrélations significatives au singulier *vs*. 6 au pluriel).
Sens au singulier est fortement corrélé à *contexte* et à *interprétation* (respectivement +0.47 et +0.41), et ses corrélats sont d'ailleurs essentiellement des candidats concepts au singulier : l'entrée est ainsi corrélée au *contexte* et non aux *contextes*, au *texte* et non aux *textes.* En outre, *sens* au singulier est négativement corrélé aux marques de formalisation (symboles, abréviations linguistiques, etc.), aux numéraux et aux parenthèses, caractéristiques des textes plus appliqués : le concept serait ainsi plus représenté dans les textes à dominante théorique.
Ces données conduisent à penser que la forme *sens* au singulier renvoie bien à une lexicalisation privilégiée de concept ; au contraire, la forme plurielle de *sens* est corrélée aux marqueurs caractéristiques des textes plus appliqués – et plus exemplifiés (pronom personnel *tu*, symboles linguistiques ?, *, ! et # et connecteurs d'exemplification). En ce sens, *sens* au pluriel serait une forme peu discutée, voire instrumentale.
On observe un phénomène similaire pour les deux formes singulier et pluriel de *corpus* – qui partagent d'ailleurs les mêmes corrélats : si on les observe aux côtés des *textes* et des *discours*, les deux formes sont corrélées à différentes caractéristiques des textes plus exemplifiés (interjections, connecteurs d'exemplification, etc.). *Corpus* serait ainsi un objet instrumental, au même titre que *sens* au pluriel.
De manière non surprenante, la *langue* et les *langues* ne renvoient pas aux mêmes objets : corrélée à la *parole* et au *langage* (substantifs et adjectifs), la *langue* semble discutée dans des textes à dominante historique, dans la mesure où elle est corrélée aux temps de l'imparfait 0.23 (et du plus-que-parfait), eux-mêmes fortement corrélés au passé simple et aux dates 0.2. *A fortiori*, elle s'oppose à de nombreuses caractéristiques des textes scientifiques (symboles, numéraux, marqueurs de structuration des textes, impératif et présent). Il en va fort différemment des *langues*, corrélées aux *langages* et au *sémantique* (substantifs et adjectifs), pour lesquelles on ne repère pas cette dimension historique.
Les corrélats morphosyntaxiques semblent ainsi représenter un critère efficace pour discriminer les candidats et les dimensions textuelles, voire même les pôles génériques (*v.* Poudat, 2006) auxquels ils sont associés (textes historiques/exemplifiés/formels, etc.) ; notons toutefois que les

[1] Fréquences absolues des substantifs *vs.* fréquences relatives des variables morphosyntaxiques.

résultats obtenus ne sont pas tous aussi probants, les candidats observés ayant des fréquences inégales. *Langue(s)* et *sens* sont ainsi particulièrement représentés et stabilisés dans le corpus.

5. Tactique

Parmi les composantes sémantiques proposées par F. Rastier, la *tactique*, qui renvoie à la *position* des unités sémantiques, intéresse particulièrement notre entreprise typologique, eu égard à la structure très normée du genre de l'article. On a ainsi apprécié la répartition des concepts dans les textes, fractionnés en dix sections de taille égale au moyen du logiciel CR développé par S. Loiseau[1], que nous appellerons *déciles de rang d'occurrences de mots par texte*.

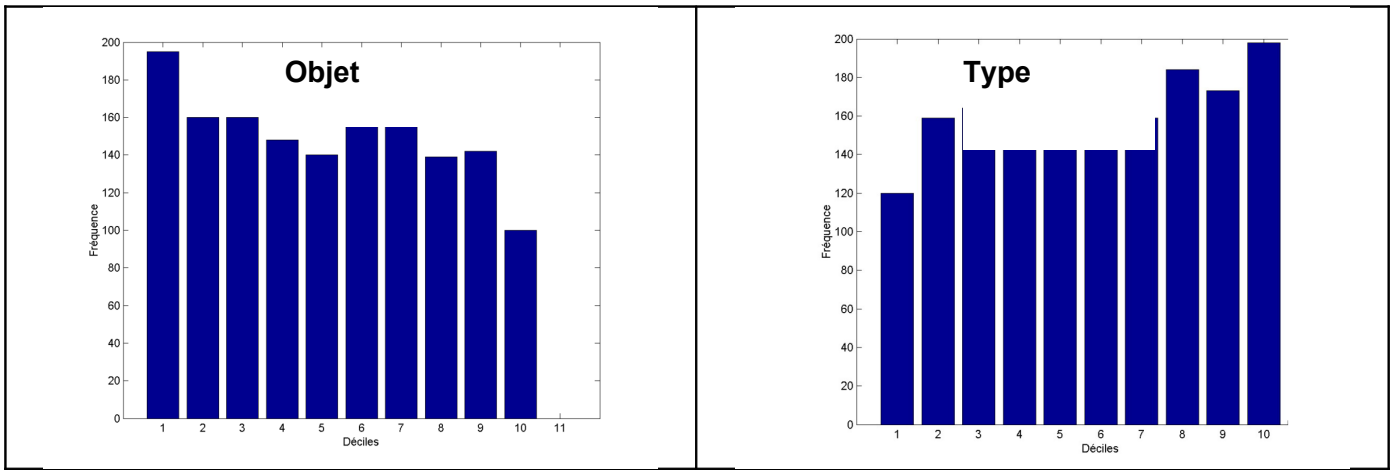
Chaque *décile* est la fréquence cumulée de l'ensemble des occurrences de l'item à cette position ; ce choix peut paraître singulier, mais Loiseau (2006) a montré que la prise en compte de la moyenne par texte (ou par unité) des occurrences à l'intérieur de chaque dixième ne modifiait pas significativement les résultats obtenus.

Cette représentation des concepts en déciles nous semble finalement plus adaptée que l'observation des candidats au sein de leurs sections textuelles, globalement très inégales[2].

Les profils obtenus sont particulièrement discriminants : certains items, comme *objet* ou *question* sont plus concentrés en début d'article et on observe une décroissance des deux entrées de l'introduction à la fin du texte. On peut légitimement penser qu'ils participent à la problématisation / exposition de la recherche présentée ; en ce sens, *objet* et *question* seraient des concepts instrumentaux plutôt que discutés.

Il en va de même des concepts de corps d'article comme *cas* ou *exemple*, de forme graduelle avec un double mouvement croissance / décroissance et un maximum obtenu en milieu d'article : peu problématisés et peu discutés en fin d'article, ces éléments sont essentiellement corrélés au développement et aux analyses menées. Ce sont donc également des concepts instrumentaux, ou méthodologiques.

Enfin, on observe des items comme *type* ou *construction*, plus denses en fin d'article : leur forme tactique est croissante, et ils semblent ainsi renvoyer aux objectifs généraux de la démarche scientifique linguistique – ici classificatoires et typologiques.



Graphique : Configurations tactiques de OBJET et TYPE (partitionnements en déciles)

Notons d'ailleurs qu'il serait intéressant de déterminer de manière plus précise et plus exhaustive les objets de début et de fin d'article, qui semblent plus méthodologiques que véritablement discutés. On pourrait ainsi les contraster d'une discipline scientifique à l'autre, afin de comparer les démarches et les présupposés méthodologiques adoptés.

Si l'on s'intéresse aux concepts discutés de l'article, la configuration tactique qui a particulièrement retenu notre attention a une forme précisément inverse de celle des concepts de corps d'article :

[1] http://panini.u-paris10.fr/~sloiseau/CR/

[2] Si l'article de revue linguistique contient en moyenne 3,46 sections de niveau 1 par texte, il n'est clairement pas soumis à la structure IMRAD (Introduction, Materials and methods, Results, Analysis, Discussion) ; en d'autres termes, son organisation est sémantiquement hétérogène (une section 2 ne renvoie à aucun objet spécifique).

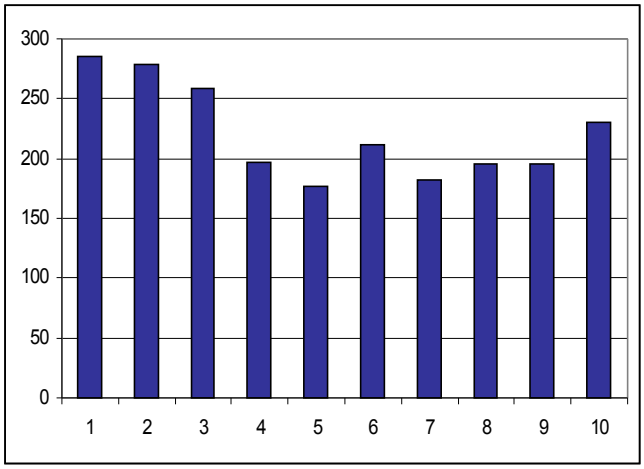


Fig. 1 : Configuration tactique de SENS

Sens est ainsi plus concentré en début et en fin d'article : on note une décroissance régulière du concept dans les trois premiers déciles, qui évoque un passage du général au spécifique (à partir du décile 4). La tendance s'inverse au-delà du décile 6, jusqu'au dixième décile – qui correspond globalement à la conclusion de l'article, où le concept revient brusquement, de manière vraisemblablement rhétorique.
Cette forme tactique incurvée semble spécifique aux concepts débattus, qui seraient ainsi davantage représentés en début et en fin d'article qu'en son corps :

Généralisation/problématisation (maximum atteint) ➔ spécification ➔ retour conclusif

Reconsidérons la liste des 20 substantifs de hautes fréquences (et les plus également répartis dans le corpus) mise au jour précédemment : *sens, forme, cas, langue, type, relation, objet, point, discours, contexte, rapport, analyse, verbe, sujet, fait, fonction, question, énoncé, phrase* et *partie*.
Outre *sens*, seuls deux candidats satisfont le critère tactique qui nous intéresse : *discours* et *langue* :

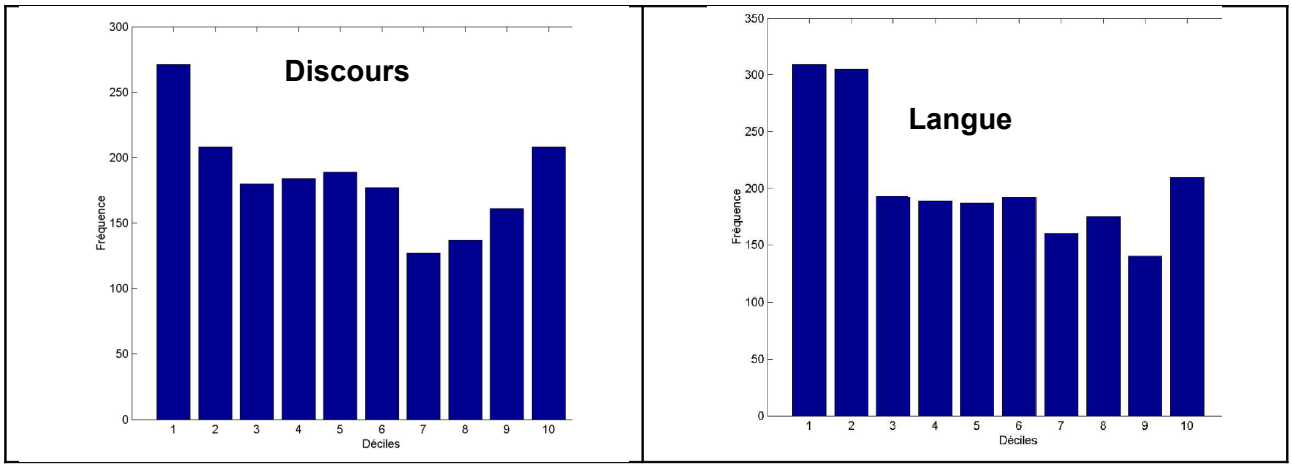


Fig. 2 : Configurations tactiques de DISCOURS et LANGUE

Malgré des différences de répartition des deux concepts dans le développement de l'article, on observe bien un retour sur le concept en fin d'article et un pic de représentation maximal en début d'article : *discours* et *langue* seraient bien des concepts de fond disciplinaire, et il en va d'ailleurs de même pour *langues* et *langage* :

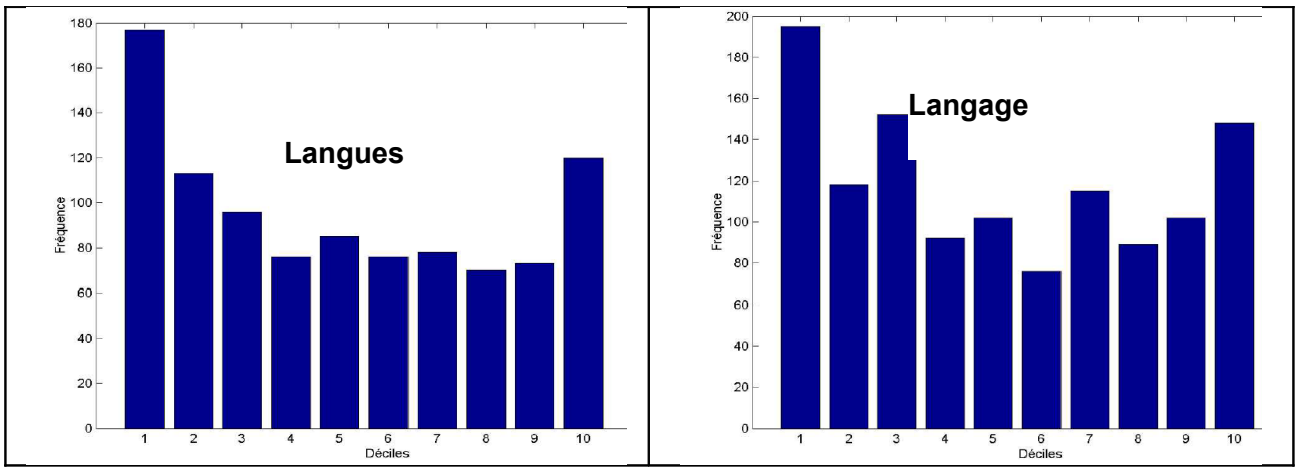


Fig. 3 : Configurations tactiques de LANGUES et LANGAGE

Si cette organisation tactique combinée au critère de fréquence nous semble ainsi permettre de discriminer les concepts discutés des autres, mentionnons une irrégularité, figurée par le substantif de haute fréquence *analyse*[1] qui manifeste également cette configuration.
Les résultats obtenus figurant davantage des formes textuelles génériques que des formes textuelles individuelles – les décomptes étant effectués au niveau du corpus -, nous nous sommes dans un deuxième temps attachée à vérifier l'existence de ces formes et leur statut d'objet discuté dans les textes du corpus.
Les configurations tactiques de chaque texte ont été observées manuellement : les résultats obtenus n'ont qu'une valeur d'approximation, l'identification des formes étant souvent délicate – il serait à terme pertinent de développer un module d'identification automatique des configurations tactiques.
Aux six concepts mis au jour avons-nous adjoint à titre illustratif quatre concepts de configuration tactique globale distincte : la forme plurielle de *discours*, *énonciation*, qui s'avère un concept de fin plutôt que de début d'article, *contexte* et *verbe*, pour lesquels nous n'avons détecté aucune forme tactique particulière.
Les résultats obtenus montrent que *langue* est le concept le plus débattu : 20% des textes qui contiennent l'entrée ont la configuration tactique qui nous intéresse, contrairement à *sens*, qui s'avère comparativement plus employé en début de texte :

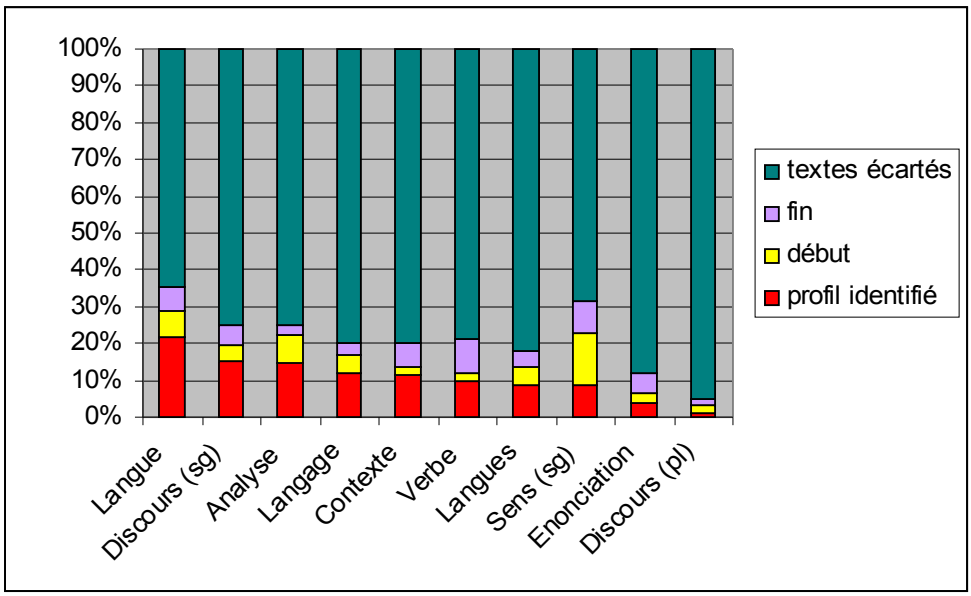


Fig. 4 : Configurations tactiques des formes les plus représentées du corpus

A fortiori, les articles ayant la configuration tactique précédemment observée semblent effectivement discuter la notion : le critère appliqué à l'objet *langage* permet par exemple

[1] Qui est néanmoins aussi un concept hjelmslevien.

d'identifier tous les textes contenant le concept dans leur titre et, de manière plus intéressante, les textes qui discutent la notion sans qu'elle soit nécessairement annoncée. Par exemple, on relève la forme tactique pour le concept de *langue* dans un article de D. Leeman[1] ; le concept est en effet discuté, ce qui n'aurait pas nécessairement été mis au jour sans ce critère.
Ce paramètre tactique, de mise en œuvre plus aisée que d'autres critères, semble ainsi particulièrement discriminant, et pourrait intéresser certaines applications de recherche d'information, en facilitant le repérage et la localisation des thèmes textuels.

6. Conclusion

Le présent article a tenté d'évaluer l'intérêt et la pertinence descriptives de plusieurs critères typologiques. L'entreprise mériterait naturellement d'être approfondie et étendue à l'ensemble des concepts du corpus d'étude, dans la mesure où nous nous sommes particulièrement intéressée aux concepts de haute fréquence *sens* et *langue*.
Si les substantifs de haute fréquence ébauchent le fond disciplinaire de la linguistique, leur seule prise en compte est insuffisante, dans la mesure où leur statut thématique demeure peu défini : s'agit-il de concepts discutés, instrumentaux ou encore des lexicalisations lieutenantes d'une forme ?
L'examen des corrélats morphosyntaxiques et lexicaux des formes permet en partie de répondre à cette interrogation en discriminant différentes acceptions du concept. Si les co-occurrents lexicaux permettent de distinguer, et d'isoler les formants, les corrélats morphosyntaxiques semblent particulièrement discriminants – à condition d'être comme nous au fait de la morphosyntaxe du corpus : ainsi, *sens* au singulier renverrait bien à une lexicalisation privilégiée de concept, tandis que *les sens* seraient des formes peu discutées, voire instrumentales car corrélées aux marqueurs de l'exemple.
La disposition tactique du candidat dans l'article permet de mettre au jour de manière très claire les formes discutées des formes non débattues, ce qui est particulièrement intéressant, dans la mesure où le critère est généralement peu pris en compte : le logiciel CR de S. Loiseau qui permet de telles manipulations de corpus nous semble ainsi particulièrement intéressant.
Si les critères pris en compte se sont avérés présenter un intérêt descriptif et discriminant, la liste est bien entendu ouverte : nous avons par exemple écarté le critère syntaxique. De surcroît, si l'on admet que les concepts les plus fréquents et les plus stabilisés font partie du fonds disciplinaire (Rastier, 2005), les concepts émergeants devraient logiquement être plus discutés, donc moins stabilisés ; il en va ainsi par exemple du concept d'*intertexte*, inconnu du glossaire bibliographique Gobert, mais de fréquence suffisante pour se prêter à l'analyse statistique – l'un des numéros thématiques du corpus[2] lui est en effet dédié :

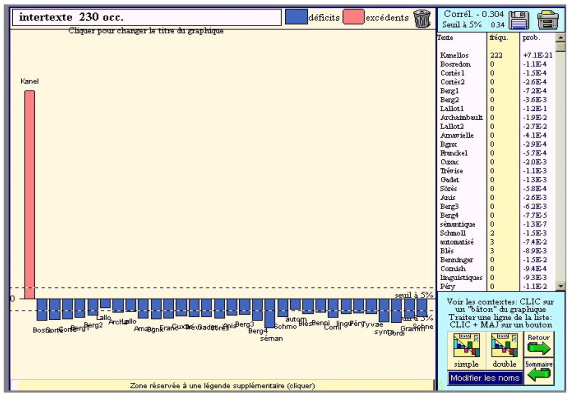


Fig. 5 : Représentation du concept d'intertexte dans les 32 numéros de revue

[1] Leeman, D. « Dans un juron, il sauta sur ses pistolets. Aspects de la polysémie de la préposition » *in* Bergounioux (*ed.*), *Approches Sémantiques des prépositions*, RSP vol. 6, Orléans, 1999.
[2] Kanellos (*ed.*), Sémantique de l'intertexte, Cahiers de Praxématique, vol. 33, Montpellier, 1999.

BIBLIOGRAPHIE

BOURION, E. 2001. *L'aide à l'interprétation des textes électronique,* Thèse, Université de Nancy II.

GOBERT, F. 2001. Glossaire bibliographique des sciences du langage, Parnormitis.

L'HOMME, M.-C. 2004. Sélection de termes dans un dictionnaire d'informatique : comparaison de corpus et critères lexico-sémantiques, in Actes. Euralex 2004, Lorient (France), 6 au 10 juillet 2004, pp. 583-593.

LOISEAU, S. 2003. Philosophical discourse from autonomy to engagement: Deleuze commentator of Spinoza, in K. Fløttum et F. Rastier (éds.), *Academic discourse — Multidisciplinary Approaches*, Oslo, Novus, pp. 36-54.

LOISEAU, S. 2005. Thématique et sémantique conceptuelle d'un concept philosophique, in G. Williams (dir.), *La linguistique de corpus*, Rennes, Presses Universitaires de Rennes.

LOISEAU, S., POUDAT, C. et ABLALI, D. 2006. Exploration contrastive de trois corpus de sciences humaines, in *Actes des 8e JADT*, 19-21 avril 2006, Besançon, pp. 631-642.

POUDAT, C. 2004. Une annotation de corpus dédiée à la caractérisation du genre de l'article scientifique, in *Workshop TCAN Construction du Savoir Scientifique dans la Langue*, Maison Alpes des Sciences Humaines, 20-21 octobre 2004.

POUDAT, C. 2006. *Etude contrastive de l'article scientifique de revue linguistique*, Thèse, Université d'Orléans.

POUDAT, C. et RINCK, F. 2006. Contrastes internes et variations stylistiques du genre de l'article scientifique en linguistique, in *Actes des 8e JADT*, 19-21 avril 2006, Besançon, pp. 785-796.

POUDAT, C. et RINCK, F. 2005. Genres scientifiques et style d'auteur : des variations stylistiques de l'article de revue linguistique, *4e Journées Internationales de la Linguistique de Corpus*, Lorient, 15-17 septembre 2005.

RASTIER, F. 2001. *Arts et sciences du texte*, Paris, PUF.

RASTIER, F. 2003. Semantics of theoretical texts in K. Fløttum et F. Rastier (éds.), *Academic Discourse, Multidisciplinary Approaches*, Oslo, Novus, pp. 15-35.

VALETTE, M. 2003. Conceptualisation and Evolution of Concepts. The example of French Linguist Gustave Guillaume, in K. Fløttum et F. Rastier (éds.), *Academic Discourse— Multidisciplinary Approaches*, Oslo, Novus, pp. 55-74.

OBSERVATIONS SUR LA NATURE ET LA FONCTION DES EMPRUNTS CONCEPTUELS EN SCIENCES DU LANGAGE

Mathieu VALETTE
ATILF CNRS

SOMMAIRE

Résumé : *On peut voir dans l'usage massif que font les sciences humaines de la métaphore et de l'emprunt aux sciences exactes au mieux une volonté d'objectivation, au pire une stratégie impressive qui parfois confinerait à l'escroquerie scientifique. Sans prétendre prendre part à cet utile débat, nous proposons ici d'étudier les procédés par lesquels le linguiste Gustave Guillaume (1883-1960) emprunta régulièrement, tout au long du processus de théorisation de sa psychomécanique du langage, des concepts ou des mots (parfois seulement leur signifiant) aux mathématiques, à la biologie et aux sciences physiques. L'étude s'inscrit dans le cadre d'une recherche sur l'analyse linguistique des textes théoriques inspirée de la philologie numérique. Parce qu'elle fait la double hypothèse qu'une théorie est un texte et que la théorisation relève de la construction du sens, cette recherche, que nous nommons épistémologie numérique, emprunte ses outils d'analyse et de description à la sémantique textuelle (F. Rastier) et ses techniques d'investigation à la linguistique de corpus. D'un point de vue méthodologique, nous nous focalisons ici sur la restitution de l'intertexte (ou intertexte simulé). Le corpus d'étude comprend le texte intégral et rédigé de 30 conférences prononcées entre 1958 et 1960. L'ensemble fait environ 100 000 mots.*

1. Introduction

1.1. Préambule

Cette étude s'inscrit dans le cadre de recherches sur l'analyse linguistique des textes théoriques menée actuellement par différents auteurs, notamment Loiseau 2005, Forest 2004, Poudat & Rinck 2006, Rastier 2005a. Pour notre part, nous faisons la double hypothèse qu'*une théorie est un texte* et qu'en conséquence, *la théorisation relève de la construction du sens*. Notre recherche, que nous avons intitulée *épistémologie numérique* (Valette 2006), emprunte ses outils d'analyse et de description à la sémantique textuelle (Rastier 2001) et ses techniques d'investigation à la linguistique de corpus.

Pratiquement, une des principales activités des scientifiques consiste à créer des concepts, les modifier, les ordonner et les articuler entre eux. Le travail de conceptualisation accompagne, voire se confond avec le travail de théorisation. Souvent victimes d'une idéologie de l'homogénéité, les ratages énonciatifs (autocorrections, lapsus, anacoluthes, etc.) apparaissent constitutifs du texte, et non seulement indices de la construction du sens, mais aussi partenaires de cette construction. Bachelard disait qu'un concept scientifique est un groupement d'« approximations successives » (Bachelard 1938, 61). Les hésitations d'un auteur, ses renoncements, ses changements d'orientation, ses palinodies, tous ces ratés de la théorisation participent à la conceptualisation et à la théorisation. Ainsi, les discontinuités du discours trouvent leur pendant dans les conditions parfois chaotiques de l'émergence des concepts, tangibles dans le texte scientifique.

1.2. L'étude

On peut voir dans l'usage massif que font les sciences humaines de la métaphore et de l'emprunt aux sciences exactes au mieux une volonté d'objectivation, au pire une stratégie impressive qui parfois confinerait à l'escroquerie scientifique. Sans prétendre prendre part à cet utile débat[1], nous proposons dans cet article d'étudier les procédés par lesquels le linguiste Gustave Guillaume (1883-1960) emprunta régulièrement, tout au long du processus de théorisation de sa psychomécanique du langage, des concepts ou des mots (parfois seulement leur signifiant) à différentes sciences.

On étudiera ici la constitution d'un thème linguistique à partir d'un fond sémantique issu des sciences exactes. En puisant presque systématiquement dans des archives étrangères à la linguistique, Guillaume tente en effet de construire ses propres concepts. L'examen des dernières conférences (années 1958-1960) montre en particulier l'appropriation par le linguiste de concepts issus de la cybernétique pour qualifier un phénomène linguistique, le passage de la langue au discours ou *actualisation*.

Notre corpus d'étude comprend le texte intégral et rédigé de 30 conférences prononcées entre 1958 et 1960. L'ensemble est lemmatisé et compte environ 100 000 mots. Pour certaines mesures contrastives, nous recourrons à un corpus de référence, composé de 340 conférences supplémentaires (textes lemmatisés, env. 1 500 000 mots) prononcées entre 1938 et 1957 et à une archive issue de la base textuelle FRANTEXT constitué de 114 essais, tous domaines confondus, publiés entre 1950 et 1960 (env. 5 710 000 mots)[2].

2. L'emprunt conceptuel : génétique du concept d'*effection* (1958-1960)

2.1. Le concept absent

Passage de la langue au discours, l'*actualisation* est réputée emblématique de la théorie de Guillaume. Pourtant, la lexie est statistiquement sous-représentée dans son œuvre. On en compte moins de 40 occurrences dans notre corpus. C'est qu'elle a revêtu plusieurs formes, qui ne correspondaient pas obligatoirement aux mêmes réalités, et qui témoignent de la part de Guillaume d'une certaine difficulté à nommer ce phénomène transitoire[3]. L'exégèse guillaumienne (par exemple Valin 1994, Joly 1987) tend en général à présenter l'*effection*, proposée en conférence par Guillaume quelques semaines avant son décès, comme un climax conceptuel qui couronnerait génialement 43 années de recherche sur l'actualisation. On ignore évidemment si Guillaume l'eût conservée et élue de la sorte s'il avait vécu quelques mois de plus. Vu sa versatilité terminologique, on pourrait légitimement en douter. Quoi qu'il en soit, les conditions de son émergence dans l'idiolecte de Guillaume méritent d'être étudiées, non pas tant en raison de son élection posthume que parce qu'il s'agit d'une des rares lexicalisations proposées. Ces conditions sont susceptibles de nous informer sur l'arrière-plan épistémologique qui lui a présidé.

2.2. La mutation

Les conférences de 1956-57 sont dominées par une problématique « hominisatrice » vraisemblablement inspirée de Teilhard de Chardin[4]. On y rencontre tout un paradigme terminologique autour du morphème *-gēn-* : *ontogénie, ontogénique, praxéogénie, praxéogénique, glossogéniques, anthropogénie, anthropogénique, endogénie, morphogénique, physiogénique*, etc. Pendant cette période marquée par une problématique évolutionniste, Guillaume, pour qualifier le processus d'actualisation, a recours au terme *mutation* : « mutation de l'indicible en dicible », « mutation de l'expérience indicible en représentation dicible », « mutation de l'expérimenté en dicible mental », etc.

Jusqu'en janvier 1959, on relève 10 occurrences de *mutation*, dont 5 en cooccurrence avec *transient*, un anglicisme que Guillaume affirme emprunter à la biologie évolutionniste qui signifie un état transitoire. Puis, c'est l'effondrement comptable. Le mot *mutation* n'est plus utilisé. Mais quelques semaines plus tard, le 30 avril 1959, apparaît un terme morphologiquement proche : *commutation* (cf. Figure 1).

---

[1] On songe aux conclusions de l'enquête de Sokal & Bricmont 1997.

[2] Cette distinction entre corpus d'étude, corpus de référence et archive a été proposée par Rastier 2005b.

[3] Cf. Valette 2006 pour une description détaillée.

[4] Les éditions du Seuil entament la publication de ses œuvres complètes à partir de 1955.

2.3. La commutation

Cette commutation, en première approximation, pourrait être associée à la notion structuraliste contemporaine. En vérité, ce n'est selon toute vraisemblance pas le cas. Le 5 mai 1959, lors de la conférence où elle sera particulièrement mentionnée (13 occurrences sur un total de 31, soit 41,9%), Guillaume définit la commutation de la façon suivante :

> Dans mon enseignement d'avant 1954-1955, […] [j'avais reconnu dans l'avance du langage en lui-même] un système de préalabilités – préalabilité du dicible par rapport au dire, préalabilité du pré-dicible par rapport au dicible […]. [Je déclare maintenant] que le langage est un système de commutations. On dira, c'est un progrès dans les mots. C'est un peu plus, parce qu'une commutation, c'est un mécanisme et qu'un mécanisme de commutation appelle une description par le dedans (Guillaume, 1995, 221)

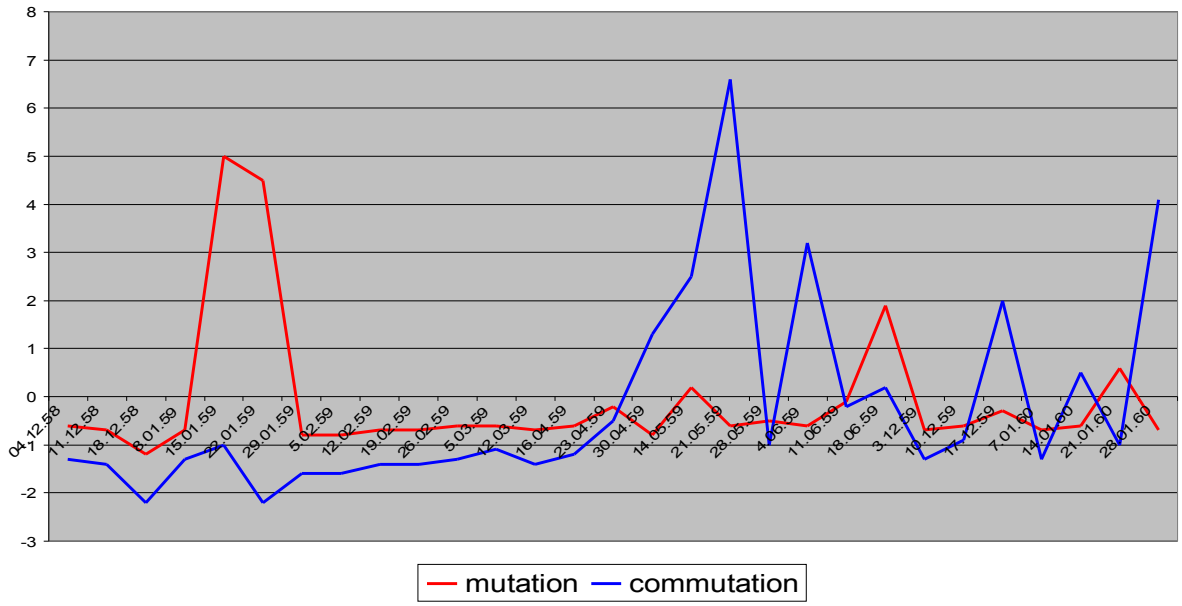


Figure 1 : distribution des mots mutation et commutation

La définition de la *commutation* comme mécanisme n'est pas inintéressante. Cette dimension est en effet absente de la *mutation*. Pour comprendre cette reformulation pour qualifier un phénomène finalement relativement constant, nous proposons de faire appel à l'intertexte, ou plutôt à *une simulation* de l'intertexte réalisée à partir de la base textuelle FRANTEXT[1]. Nous avons sélectionné une archive composée de 114 essais publiés entre 1950 et 1960, comprenant le mot *commutation*, c'est-à-dire susceptibles de restituer les usages de ce mot à l'époque où Guillaume choisit de se l'approprier. Nous avons ensuite utilisé la fonction de comptage « voisinage d'un mot » proposée par l'interface d'accès à FRANTEXT. Nous l'avons paramétrée de façon à obtenir la liste des cooccurrents dans une fenêtre de 20 mots (cf. Figure 2).

[1] http://www.atilf.fr/frantext.htm

20 commutation	2 calculateur	2 précisément
6 relations	2 canoniques	2 relation
5 Heisenberg	2 charme	2 relie
5 téléphoniques	2 circuits	2 simple
4 équations	2 communication	2 studio
3 cas	2 commuter	2 théorème
3 jonction	2 corps	2 thèse
3 mécanique	2 câbles	1 Bell
3 problèmes	2 groupe	1 Bohr
3 radiodiffusion	2 liaisons	1 Cdm
2 Schur	2 modulation	1
2 appareils	2 opérateurs	Vertauchungsrelation
2 brancher	2 problème	1 Wedderburn
		[…]

Figure 2 : Voisinage FRANTEXT (ATILF) dans 114 essais publiés entre 1950 et 1960 comprenant le mot commutation, fenêtre de 20 mots (extrait)

D'emblée, on devine quel est l'usage privilégié qui est fait du mot *commutation* à cette époque-là, il s'agit indubitablement d'un terme issu des mathématiques – et plus particulièrement de la théorie de l'information. Mais si Guillaume l'emprunte à cette science, en fait-il le même usage ? Indubitablement pas. Si nous nous intéressons aux spécificités de la conférence du 21 mai 1959, nous constatons que les domaines caractéristiques qui y sont actualisés, par rapport à l'ensemble de notre corpus d'étude (1958-1960), sont liés à la neurologie et non aux mathématiques ou à la théorie de l'information (Figure 3).

Forme	Frq. Tot.	Fréquence	Coeff.
vu_en_pensée	39	18	18
dicible	44	15	13
commutation	31	13	13
grammairien	27	11	11
pédagogue	6	6	10
unité	23	8	8
hypobasique	7	5	7
grammaire	94	14	7
tardif	15	6	7
connaissance	54	10	7
médecin	4	4	7
isologie	4	4	7
phrase	69	12	7
neurochirurgien	3	3	6
neurophysiologue	5	4	6

Figure 3 : Spécificités (lemmes) de la conférence du 21 mai 1959 par rapport au corpus d'étude (logiciel Lexico3, Paris 3)

Mieux encore, il apparaît que le préfixe *neuro-* et le mot *commutation* sont les deux éléments les plus caractéristiques de cette conférence comparée à l'*ensemble* des 370 conférences qui constituent notre corpus de référence (Figure 4).

Forme	Frq. Tot.	Fréquence	Coeff.
neuro-	13	11	29
commutation	31	13	28

Figure 4 : Spécificités de la conférence du 21 mai 1959 par rapport au corpus de référence composé des conférences 1938-1960 (logiciel Lexico3, Paris 3)

Ces quelques données montrent à quel point le concept de commutation semble associé, dans la pensée de Guillaume, au domaine de la neurologie, *à l'exception de tout autre*, lui-même absolument spécifique à cette leçon. Le neurologique fait en effet une entrée en force au moment où *commutation* est le plus commenté.

2.4. Restitution des thèmes informulés

Que s'est-il passé entre l'abandon du terme *mutation* et l'adoption de *commutation* ? Guillaume, pendant cette période, semble s'intéresser à des problèmes qu'il traite peu ailleurs, tels que ceux, très singuliers de l'« imagination constructive » d'Antoine Meillet et du « rêve constructif » d'Henri Poincaré[1]. Corrélativement à ce dernier, ce sont également les mathématiques qui sont questionnées de façon peu commune. À la vérité, si l'on excepte quelques œuvres de jeunesse (publiées entre 1911 et 1913 et jamais rééditées), Guillaume ne traite jamais autant, sur un laps de temps aussi court, des mathématiques qu'en 1956, lors d'une conférence entièrement consacrée à la cybernétique.

Nous avons montré ailleurs (Valette 2006 et *à paraître*, 117-136) que la cybernétique constituait un thème peu explicité mais néanmoins important dans le projet théorique de Guillaume, qui sans doute considérait sa « psychomécanique du langage » comme une forme de cybernétique. Ces observations reposent sur un corpus positif (i.e. où la cybernétique est explicitement mentionnée) essentiellement composé de conférences et d'essais rédigés en 1956. Dans nos précédentes publications (ainsi que dans l'*à paraître*), nous nous limitions à ce corpus positif. Nous sommes maintenant en mesure d'affirmer que ce thème, apparemment absent des conférences et articles ultérieurs à 1956, fait en réalité un retour très significatif en 1959, mais sans être explicité.

On relève en effet, dans l'année universitaire 1958-1959, trois passages intéressants annonçant cette prégnance de la cybernétique dans les dernières conférences de Guillaume. Le premier se situe au tout début de l'année, le 4 décembre 1958. Guillaume évoque alors – sur un ton peut-être las, peut-être ironique – non pas la cybernétique, mais « des cybernéticiens », à propos d'une trinité pensante : l'homme, l'animal, la machine. Puis, de façon plus implicite, le 5 mars 1959, il est question de la langue comme d'un « dispositif mécanique d'inclusion » (Guillaume 1995, 162). Actualisée lors de la conférence où le domaine des mathématiques est le plus saillant, l'expression est loin d'être anodine parce qu'elle s'insère dans une discussion sur la dialectique de la liberté et de la contrainte, de l'ordre et du désordre. La langue y est en quelque sorte présentée comme un homéostat.

[1] Sur ces différents thèmes, on pourra consulter Valette, à paraître.

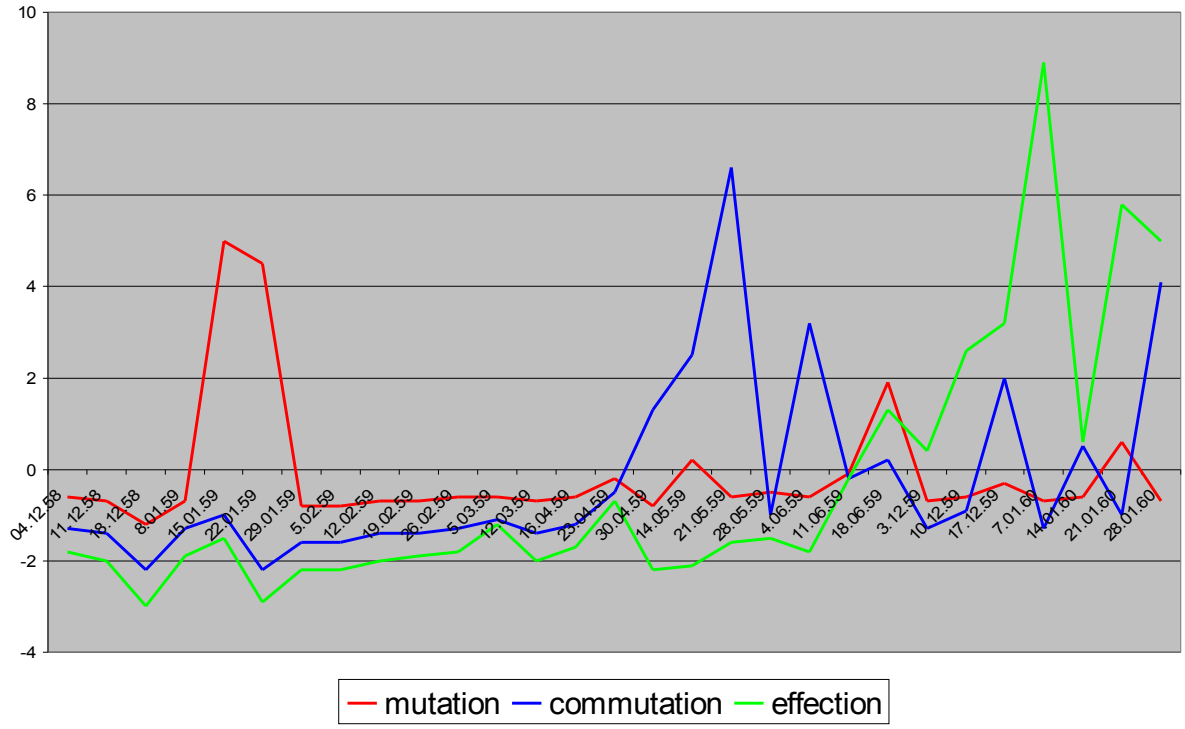


Figure 5 : distribution des mots mutation, commutation et effection

Le troisième passage distingué méritera toute notre attention car il présente un intérêt philologique particulier : le 14 mai 1959, alors que Guillaume est en pleine élaboration du thème corrélant la commutation et le domaine de la neurologie, il évoque « la partie mécanisable de la pensée ». L'expression pourrait paraître quelconque – et très guillaumienne dans l'esprit – mais « mécanisable » est, dans notre corpus de référence, un hapax : il n'en existe aucune autre occurrence, dans l'œuvre publiée tout au moins. C'est pour nous l'indice de son importance[1] . Malgré une racine familière, cet adjectif néologique n'appartient pas à l'idiolecte de Guillaume. Qu'il y ait recours trahit ses lectures : toutes les occurrences relevées sur FRANTEXT sont extraites de textes consacrés à la cybernétique[2], et notamment de l'ouvrage *Les machines à penser*, de Louis Couffignal, publié en 1952. Guillaume l'a sans doute lu aux environs de 1955, comme l'atteste la mention de telles « machines à penser » dans les notes des conférences inédites du 6 janvier 1955 (f°11) et du 9 juin 1955 (f°17-20).
Ainsi, nous pouvons faire l'hypothèse que dans le courant de l'année universitaire 1958-1959, Guillaume, sans vraiment le dire explicitement, rouvre le dossier cybernétique délaissé quelques années auparavant. Ses lectures et ses notes de lectures lui permettent d'élaborer le thème commutation/neurologie et, subséquemment, le fameux concept d'*effection*.

2.5. De la commutation à l'effection

Malgré sa fortune dans les dernières semaines d'enseignement de l'année universitaire 1958-1959, la commutation cède la place à la rentrée 1959-1960 au concept plus restreint d'*effection*, lui-même défini le 10 décembre 1959 par rapport à celle-ci dans les termes suivants : « *effection* voulant dire commutation de la puissance en effet » (Guillaume, 1995, 262, cf. Figure 5). Ainsi, le thème du passage et de la transition – de la langue au discours, du dicible au dire, de la puissance à l'effet, etc. –, thème de l'actualisation innommée en somme, trouve une ultime formulation, morphologiquement simple et là encore, parfaitement adaptée au vocabulaire métalinguistique de Guillaume : il lui aura suffit de dériver le terme *effet*.

[1] À rebours du dogme textométrique, nous considérons en effet qu'en matière de construction textuelle des concepts, la rareté fait parfois la valeur.
[2] À l'exception d'une, issue de *Mathématiques*, J. Roubaud, 1997, Seuil.

7 effection	2 set	1 automates
5 organes	2 système	1 celles
4 champ	2 thèse	1 central
2 caractère	2 équivalent	1 confirmer
2 comportement	1 Lorente de No	1 constituante
2 fibres	1 acte	1 contrôler
2 géographique	1 action	1 cortex
2 machines	1 aperception	1 couplées
2 monde	1 appel	1 cybernéticiens
2 musculaire	1 asservis	1 cérébral
2 selon	1 attraction	[…]

Figure 6 : Voisinage FRANTEXT (ATILF) dans 114 essais publiés entre 1950 et 1960 comprenant le mot effection, fenêtre d'une phrase (extrait)

Toutefois, là encore, l'effection est loin d'être immanente à l'appareil théorique ; l'analyse de notre archive, c'est-à-dire de l'intertexte simulé à partir d'une sélection d'essais issus de FRANTEXT (cf. supra, 3.3), montre que les cooccurrents d'*effection* ont trait à la fois aux machines cybernétiques *et* au cerveau (cf. Figure 6). Autrement dit, *effection* constitue, en quelque sorte, une *contraction thématique*, ou la synthèse conceptuelle du thème associant d'une part, les domaines Théorie de l'information et Mathématiques (lexicalisés par *commutation*) et d'autre part, le domaine neurologie.

3. Conclusion : simuler l'intertexte ?

L'« épistémologie numérique » vise à objectiver l'étude des théories scientifiques en substituant aux connaissances encyclopédiques et aux intuitions afférentes de l'épistémologue, une instrumentation reposant pour partie sur les statistiques textuelles et la linguistique de corpus. Objectiver implique donc de prendre en compte la construction et la circonscription des observables. Confronté à la question des frontières du corpus[1], nous avons fait, dans Valette 2006, la proposition méthodologique suivante : le texte d'un auteur constitue une unité en soi et il ne nécessite pas de recourir à d'autres sources pour l'expliquer, sauf lorsque celles-ci sont positivement mentionnées dans le corpus initial. Cette position d'inspiration structuraliste visait tout particulièrement à se préserver des explications invoquant le *contexte* psychologique, sociologique et historique de la production scientifique. En choisissant de traiter de l'emprunt des concepts dans le présent papier, nous avons souhaité éprouver notre position initiale : il semble en effet improbable d'étudier les migrations conceptuelles en ignorant les sciences sources et en se cantonnant à l'analyse d'une théorie cible. C'est ainsi que nous avons « ouvert » notre corpus. Il n'a cependant pas été question de l'ouvrir au contexte socio-historique – *sur lequel le linguiste n'a, de notre point de vue, rien à dire* – mais à un *intertexte simulé.*

Nous avons en effet recouru à la base textuelle FRANTEXT pour restituer l'environnement textuel dans lequel Guillaume aurait hypothétiquement puisé son inspiration. Bien qu'on puisse difficilement parler de lexicométrie à propos de la fonction « voisinage d'un mot » que nous avons utilisée, dans la mesure où celle-ci compte mais ne pondère pas, il nous semble que l'analyse atactique des seuls cooccurrents (i.e. sans que soient pris en compte les énoncés cotextuels proprement dits ni d'éventuelles définitions subséquentes) donne à entrevoir la texture sémantique (isotopies et thèmes sémantiques) d'un intertexte. Qu'*effection*, au sens cybernétique, signifie « Action de répondre à un stimulus » (d'après le *Trésor de la Langue Française*, ou *TLF*) nous importe moins que d'observer que les mots *organes*, *cortex*, *machines* et *automates* se rencontrent dans son voisinage. La définition d'*effection* par Guillaume, « commutation de la puissance en effet », n'a aucun rapport avec celle qu'en donne le *TLF* ; elle n'est en revanche peut-être pas étrangère à cet entour sémantique fait à la fois de cellules grises et de rouages.

[1] Question non rhétorique qui nous a effectivement été posée par Arild Utaker, à l'issue d'un exposé présenté à Paris en février 2003 dans le cadre du *KIAP Project* (Université de Bergen). Le texte de cet exposé a été publié dans Fløttum & Rastier, éds. 2003.

BIBLIOGRAPHIE

Ouvrages théoriques

BACHELARD, G. 1938. *La formation de l'esprit scientifique. Contribution à une psychanalyse de la connaissance objective*, Paris, Vrin, 13ème édition en 1986.

FOREST, D. et MEUNIER, J.-G. 2004. Classification et catégorisation automatiques : application à l'analyse thématique des données textuelles, in G. Purnelle, C. Fairon et A. Dister (dir.), *Le poids des mots. Actes des 7ièmes Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, Presses Universitaires de l'Université Catholique de Louvain, Volume 1, pp. 434-444.

JOLY, A. 1987. *Essais de systématique énonciative*, Lille, Presses Universitaires de Lille.

FLØTTUM, K. et RASTIER, F. (éds.) 2003. *Academic discourse. Multidisciplinary approaches*, Oslo, Novus Press.

LOISEAU, S. 2005. Thématique et sémantique contextuelle d'un concept philosophique, in G. Williams (éd.), *La Linguistique de corpus. Actes des deuxièmes journées de la linguistique de corpus*, Rennes, Presses Universitaires de Rennes, pp. 129-140.

POUDAT, C., RINCK, F. 2006. Contrastes internes et variations stylistiques du genre de l'article scientifique de linguistique, in J.-M. Viprey (dir.), *Actes des 8è journées internationales d'analyse statiques des données textuelles*, pp. 785-796.

RASTIER, F. 2001. *Arts et sciences du texte*, Paris, PUF.

RASTIER, F. 2005a. Pour une sémantique des textes théoriques, *Revue de sémantique et de pragmatique*, 17, pp. 151-180 ; publié dans la revue électronique *Texto ! Textes et cultures,* rubrique Dits et inédits.

RASTIER, F. 2005b. Enjeux épistémologiques de la linguistique de corpus, in G. Williams (éd.), *La linguistique de corpus*, Rennes, PUR, pp. 31-45 ; publié dans la revue électronique *Texto ! Textes et cultures,* rubrique Dits et inédits.

SOKAL, A., BRICMONT, J. 1997. *Impostures intellectuelles*, Paris, Odile Jacob.

VALETTE, M. 2006. La genèse textuelle des concepts scientifiques. Étude sémantique sur l'œuvre du linguiste Gustave Guillaume, *Cahiers de Lexicologie, 2/2006* ; prépublié dans la revue électronique *Texto ! Textes et cultures,* rubrique Dits et inédits.

VALETTE, M. (en cours de publication). *Linguistiques énonciatives et cognitives françaises. Gustave Guillaume*, *Bernard Pottier*, *Maurice Toussaint*, *Antoine Culioli*, collection « Bibliothèque de Grammaire et de Linguistique », Paris, Honoré Champion.

VALIN, R. 1994. *L'envers des mots. Analyse psychomécanique du langage*, Québec, Presses de l'Université Laval / Paris, Klincksieck.

Corpus (références des éditions papier)

COUFFIGNAL, L. 1952. *Les machines à penser*, Paris, Editions de Minuit.

GUILLAUME, G. 1971-2001. *Leçons de linguistique 1938-1960*, 16 volumes, Québec, Presses de l'Université Laval ; Paris, Klincksieck et Lille, Presses Universitaires de Lille.

GUILLAUME, G. *Manuscrits et mémoires,* Fonds Gustave Guillaume, Québec, Université Laval.

ÉCONOMIE CINÉTIQUE ET FORMES DE MIMESIS : LE CAS DES HISTOIRES DE VIE

Jean-Michel BAUDOUIN & Juan PITA
Faculté de psychologie et des sciences de l'éducation
Université de Genève

SOMMAIRE

Résumé : *Dès lors que l'on identifie ce que l'appréhension de l'action doit au texte, et ceci de manière principielle, et dès lors que l'on fait l'hypothèse que le plan majeur de structuration textuelle est défini par la dimension du genre de texte, on est conduit à thématiser ce qu'une herméneutique de l'action, nécessairement diverse et multiple, doit aux genres de textes qui configurent cette action (Baudouin, 2004).*

Ceci est le cadre général d'une problématique de l'action, où l'analyse et la conceptualisation du genre textuel prennent une dimension doublement stratégique (Baudouin, 2005). (i) Le genre, par la médiation des structurations textuelles qu'il constitue, et en conséquence les configurations changeantes de l'action qu'il entraîne, représente un plan opératoire essentiel de l'analyse : les genres de textes déterminent la configuration de l'action. (ii) Les genres de textes sont cependant sous la dépendance des cours d'action et des pratiques qui les accueillent ou les suscitent : ils relèvent pleinement d'une problématique praxéologique dont il convient d'examiner les divers régimes possibles de réalisation.

Afin d'éviter un tour trop général à notre propos, nous appuierons nos propositions sur quelques résultats d'une analyse ayant porté sur des récits autobiographiques, récits réalisés par des étudiants dans le cadre d'un séminaire en sciences de l'éducation à l'Université de Genève (Baudouin, 2001). En relation avec les deux orientations précédemment retenues, nous montrerons (i) en quoi l'action des protagonistes, et plus largement les « matériaux biographiques » livrés par le récit, sont configurés par l'opération narrative, telle que celle-ci est elle-même retravaillée spécifiquement par le champ générique de l'autobiographie. Nous indiquerons en particulier les formes de mimèsis qui sont intégrées à son fonctionnement propre, et qui définissent une extrême sélectivité dans les actions retenues par le récit, voire les périodes biographiques privilégiées. (ii) Nous montrerons en quoi les manières distinctes prises par l'autobiographie dans le corpus d'analyse (récit autobiographique versus autoportrait) requièrent la prise en compte du contexte d'effectuation (le séminaire) et des modèles textuels qui dans le cas présent y sont en compétition. (iii) Nous développerons enfin en quoi la numérisation du corpus permet de repérer et d'analyser une économie cinétique des récits.

1. Le contexte d'effectuation

Les sciences de l'éducation abritent depuis les années 80 un champ de recherche propre à la formation des adultes, repéré sous la dénomination histoire de vie (Le Grand & Pineau, 1993) et dédié à une approche autobiographique des processus de formation des adultes : ceux-ci sont invités, dans le cadre d'un séminaire de formation, à produire un « récit de vie » privilégiant les dimensions éducatives de leur histoire (Lainé, 1998). Signalons en quelques mots que ce champ dispose d'une association internationale, de réseaux de formation et de recherche assurant l'animation locale et d'un ensemble de publications faisant le point sur les travaux réalisés. Il contribue, au sein des sciences de l'éducation, à nourrir les analyses propres à l'éducation informelle et à l'autodidaxie. Il a orienté les travaux préludant aux dispositifs légaux de reconnaissance et de validation des acquis de l'expérience. Il comporte une dimension militante de

promotion sociale : les histoires de vie en formation sont également des dispositifs de formation visant à favoriser l'intégration des adultes en des cursus de formation professionnelle ou universitaire, pour lesquels ils s'estiment, à tort ou à raison, peu légitimés.
On peut faire l'hypothèse que ce recours à l'approche autobiographique des processus de formation n'est pas sans lien avec un phénomène plus général, de « démocratisation » de l'autobiographie. On peut l'observer par exemple en sociologie, donnant une dignité épistémique à la prise en compte biographique des sujets sociaux d'origine modeste, à la mise en forme de leurs parcours de vie et à l'identification de leurs cultures propres (Demazière & Dubar, 1997). Un phénomène d'allure similaire est repérable au plan éditorial, où le genre autobiographique connaît un développement qui ne semble pas désemparer (Chiss, 1985), avec de remarquables prolongements au plan de l'étude dans le champ de la critique littéraire.
En ce qui concerne Genève, ces séminaires se déroulent dans le cadre de la Licence en sciences de l'éducation depuis le début des années 80 (pour une analyse systématique : Baudouin, 2001). Ils comportent une trentaine de séances hebdomadaires ventilées sur l'ensemble de l'année académique et présentent quatre étapes : une phase d'enseignement centré sur les histoires de vie dans le champ des sciences humaines et sociales, une phase de récit autobiographique oral dans le cadre d'un groupe restreint, une phase de production du récit écrit, une phase d'analyse de ces récits écrits. Le moment de rédaction se situe à mi-parcours de l'année, durant la période de pause de quelques semaines située en février-mars, période où les effectifs sont dispersés. Les participants n'ignorent pas que le texte produit sera lu par l'ensemble des membres du groupe restreint[1] et servira de base à une réflexion ultérieure[2]. Aucun des écrits produits dans ces séminaires ne fait l'objet d'une évaluation sommative. Par contre, les participants savent en s'engageant qu'ils auront à produire un écrit à caractère autobiographique, et que cette production est nécessaire pour « valider » le séminaire dans le cursus de licence. Le dispositif ainsi décrit peut concerner selon les années de un à quatre groupes restreints, travaillant en parallèle avec leurs responsables spécifiques (600 textes environ ont été ainsi produits en quelques 25 années)[3].
Du point de vue de la problématique des genres de texte, observons que nous sommes assez loin de la posture constituant un champ générique par identification et discussion de propriétés caractérisantes, et légitimant *a posteriori* une recollection. Notre corpus ne tient que par le contexte. C'est-à-dire un séminaire universitaire optionnel, au cours duquel des étudiants rédigent un récit à caractère autobiographique, pour lequel les « consignes » initiales sont plutôt vagues. Notre corpus ne méconnaît cependant pas totalement les « lois du genre » : mais l'identification de ce qui est requis comme production textuelle porte davantage sur un champ générique, c'est-à-dire une catégorie plus large du type *écrit à caractère autobiographique*, que sur un genre de texte clairement codifié et identifié[4]. Nous pouvons donc nous attendre à une variabilité liée d'une part à l'indétermination relative propre à une catégorie large permettant différentes spécifications et d'autre part à la *représentation* que l'auteur potentiel se fait du genre de texte à produire. *Nous disposons donc et pour le moins de deux principes possibles de variation à l'œuvre dans le corpus : une variation textuelle propre à l'empan générique et une variation liée au sujet et à la représentation qu'il se fait du texte à produire.* Observons que le second principe de variation est

[1] Huit à neuf participants en général. Le corpus étudié comporte 22 récits analysés qui proviennent de trois séminaires différents animés par Jean-Michel Baudouin, en 1990, 1995 et 2000.
[2] Le public étudiant fréquentant la Licence en sciences de l'éducation comporte deux grandes catégories : des jeunes gens en formation initiale, provenant de l'enseignement secondaire et des professionnels de l'enseignement, de la santé et du travail social le plus souvent, qui viennent au titre de la formation continue.
[3] Après demande d'autorisation auprès des auteurs, le corpus fait l'objet d'une mise au point particulière, permise par la numérisation des textes. Il est procédé aux adaptations mineures requises par la garantie d'un anonymat, aussi bien pour les auteurs que pour les tiers évoqués (essentiellement en modifiant les noms et prénoms des auteurs, les prénoms des tiers, les « raisons sociales » des institutions, les toponymes correspondant à des localités de taille petite ou moyenne, ainsi que les dates trop précises du type le *24 janvier 1974*). Ce travail ne touche pas au « corps » du texte, qui est rigoureusement préservé, coquilles originales comprises.
[4] Il existe un autre type d'indétermination, propre à la multiplicité des appellations dans le champ scientifique concerné. Les corpus propres à l'approche biographique en sciences sociales sont désignés diversement sous la plume des chercheurs : autobiographies, biographies éducatives, histoires de vie, récits de formation, récits de pratique, récits de vie. Ils fonctionnent en fait en variation libre, y compris dans les ouvrages spécialisés (par exemple Poirier, Clapier-Valladon & Raybaut, 1983 ou Penef, 1990).

une *constante* de toute production textuelle, qui repose toujours sur une série de compétences propre au sujet.

2. La problématique des genres

La discussion que nous allons présenter ici s'appuie sur les travaux de Bakhtine (1984) et de Rastier (2001) d'une part (pour la problématique générale des genres) et sur les travaux de Beaujour (1980) et de Lejeune (1975, 1996). Nous reprenons à notre compte la perspective ouverte par Bakhtine, selon laquelle toute production suppose l'appui sur des mises en forme déjà présentes dans les divers sites de l'activité humaine. Nous verrons que dans le cas de notre corpus, deux genres semblent en compétition. Rappelons que l'approche de Bakhtine suppose un achèvement de l'énoncé dans une *totalité*, qui ne se résume pas à son intelligibilité linguistique, mais est sous la dépendance de trois facteurs liés dans l'énoncé : « 1) l'exhaustivité de l'objet de sens, 2) le dessein, le vouloir-dire du locuteur, 3) les formes-types de structuration du genre de l'achèvement » (1984, p. 282). Dans le cas de notre corpus d'étude, la totalité visée est comprise comme *l'étendue* d'une vie et définit la culture la mieux partagée parmi nos auteurs[1], aux côtés d'un *vouloir dire* privilégiant les dimensions de l'histoire personnelle, ceci en adéquation avec les « lois » plus générales du genre de l'autobiographie. Cette hypothèse d'une culture communément partagée concernant l'autobiographie est également proposée par Lejeune, lorsqu'il aborde (1996, p. 51) l'analyse des *Carnets* non publiés de Sartre, où il voit le prototype de toute entreprise autobiographique :

> L'intérêt est de voir comment Sartre s'y prend pour composer un texte autobiographique. Il a écrit des textes théoriques sur la *méthode* biographique. Comment fait-il quand il est au pied du mur, et qu'il s'agit de lui ?
>
> Et il ne s'agit pas seulement de Sartre. Peut-être ses comportements nous apprendront-ils quelque chose sur la situation où se trouve n'importe quelle personne qui entreprend d'écrire sa vie.

Parmi ces « comportements » les plus récurrents figure l'identification de permanences : « La recherche de la constante est une des constantes de la recherche [au sens de préoccupation ou de projet] autobiographique » énonce Lejeune, (*op. cit.*, p. 65), orientation proposée comme axiome du genre : le récit autobiographique se construit sur la base de régularités dont on recherche les origines dans l'enfance. On tiendrait ici la rhétorique élémentaire de toute entreprise autobiographique, conduisant les auteurs à l'introduction d'un ordre dans le flux biographique : comment raconter sa vie, par où commencer, et surtout comment choisir parmi la quantité des « choses » pouvant peupler une vie ?

3. Deux genres en présence

Dans notre corpus, la nécessité d'introduire une sélection pour rendre praticable le projet autobiographique se fait selon deux orientations distinctes : soit il y a choix d'une périodisation chronologique, où chaque séquence sera ensuite animée par quelques événements estimés importants, sur la base de catégories biographiques élémentaires[2] (« l'école », « les vacances », « la famille », « le travail », etc.), soit il y a recours à une problématique ou un intérêt de connaissance particulier (par exemple le féminisme, la psychanalyse, la « figure des maîtres »), le texte ne reposant plus uniquement sur une perspective chronologique, qui est même en certains cas complètement abandonnée, pour adopter des logiques argumentatives où les évocations autobiographiques sont convoquées à titre illustratif.

Dans le premier cas, nous avons des textes à dominante narrative (voir extrait 1 en fin de contribution) ; dans le second cas, nous avons des textes à dominante argumentative (voir extrait 2).

Lorsque l'on se tourne du côté de la littérature critique (Beaujour, *op. cit.* ; Lejeune, *op. cit.*), on ne rencontre guère de difficulté à identifier ces deux extraits comme typiques de « l'écriture du soi »

[1] Nous disposons cependant d'un texte qui déroge à cette totalité et qui s'attache à narrer par le menu la 26ème année de son auteur : mais celui-ci argumente et légitime longuement ce choix en introduction de son écrit, probablement parce qu'il a conscience de réaliser un écart par rapport à une attente de ses destinataires (les autres membres du séminaire). De manière corollaire, les auteurs qui réalisent la totalité visée de l'étendue d'une vie ne justifient ni n'argumentent leurs choix. Ainsi la présence d'un récit d'enfance semble une des caractéristiques les mieux reçues de l'autobiographie.

[2] Genette propose la notion de syllepses pour ces opérateurs sémantiques permettant de regrouper des aspects itératifs typiques de l'autobiographie (Genette, 1983). Nous y reviendrons.

prise comme champ générique (identité des figures de l'auteur, du narrateur et du personnage principal ; visée référentielle), et à ranger le premier extrait dans la catégorie de l'autobiographie (visée rétrospective et chronologique, avec dominante narrative centrées sur l'évocation d'une vie) par opposition au second extrait, typique de l'autoportrait (visée argumentative à dominante discursive dédiée à une présentation de soi). Selon les travaux de Beaujour, dans l'autoportrait :

> le *je* résume la structure du monde, comme le microcosme celle du macrocosme. Par suite, le discours de *je* et sur *je* devient un microcosme du discours collectif sur l'univers des choses – *chose* étant pris au sens de *res* : sujet à traiter, lieu commun, *topos* (*ibid.*, p. 30).

C'est dans le genre du *speculum* de la littérature médiévale que Beaujour propose d'identifier le type de référence, en tant qu'il constitue un *rassemblement encyclopédique de connaissance*. L'autoportrait procède dans une telle orientation d'un « miroir d'encre », entre le sujet et le monde, le *je* et l'univers. Il définit ainsi une adaptation du *speculum* du Moyen Age, en particulier parce que son trait distinctif n'est plus narratif comme dans le récit, mais *topique*. L'abandon de la restitution d'un *je* passé au profit d'un *je* présent s'accompagne de l'émergence d'un vis-à-vis constitué par le monde externe : traiter ce monde, c'est adopter les *topos* du *speculum*, c'est-à-dire les savoirs d'une *doxa*.

Dans notre corpus d'étude, quatre textes correspondent très clairement à l'autoportrait, alors que dix-huit relèvent pleinement du récit autobiographique. Notre problème est alors de comprendre comment opère la mise en œuvre d'un genre de texte. Dans le cas du récit autobiographique, la vogue éditoriale de l'autobiographie permet d'envisager avec un réalisme acceptable sa réception auprès du plus grand nombre et en conséquence un horizon d'attente partagé. Les dix-huit récits de notre corpus présentent cet « allant de soi » non problématique. Comme nous l'avons vu, il faut qu'il y ait dérogation à une loi du genre attendue (quand la totalité visée ne porte que sur une seule période biographique au lieu de l'entier du temps de vie) pour qu'il y ait alors discussion et justification argumentée du choix opéré. Mais le recours à l'autoportrait attesté dans quatre cas nous prive de l'hypothèse commode de la réduplication d'une forme faisant partie d'un horizon d'attente partagé. Il y a rupture dans la réception du genre : on peut imaginer nos auteurs lecteurs de Cavana ou Annie Duperey, mais de Montaigne ou Leiris, c'est moins probable. Ces quatre textes se caractérisent par des traits saillants et communs (problématisation, références théoriques, prépondérance du narrateur, dislocation ou sélectivité du narratif), lesquels recoupent les dimensions mises en avant par les travaux de Beaujour : indexation des thématiques sur les « rubriques et divisions » d'une culture, travaillées par les idéologies et les sciences d'un moment historique particulier. L'autoportrait est à prendre comme une destitution de la catégorie du récit dans la perspective d'une saisie de soi par le sujet : l'intrigue ne suffit pas (ou plus) à une auto-compréhension, qui requiert alors la mobilisation des lieux communs du savoir vulgarisé. L'autoportrait, comme le *speculum*, est une médiation didactique entre le manuel systématique détaillant les savoirs de référence concernés et une écriture du soi (voir extrait 3).

4. Le genre comme résultante d'une épreuve

Certes on peut faire l'hypothèse que nous sommes ici très nettement du côté du « texte académique », typique des *opus* produits dans le cadre d'une formation supérieure universitaire. Ces quatre textes reposeraient sur un mode de génération qui est celui des travaux universitaires, de taille moyenne, requis pour « valider » les cours d'un cursus de Licence en contrôle continu. L'autoportrait serait donc dans notre cas un genre composite, se définissant comme une sorte de transaction entre deux genres, celui du récit autobiographique et celui du « texte didactique » (écrits rédigés par les étudiants dans un cadre évaluatif). On peut ajouter qu'il n'est pas surprenant que, dans le cadre d'un séminaire universitaire, le genre de l'écrit « correspondant » au contexte reprenne ses droits. Certes. Mais un tel propos supposerait symétriquement que les auteurs rédigeant une autobiographie aient conscience que ce genre de texte introduise un écart par rapport au genre de l'autoportrait et une norme prescriptive liée au contexte académique, écart nécessitant alors une justification ou un argumentaire. Or, force est de constater que c'est le contraire qui se produit. Ce sont les auteurs de l'autoportrait qui explicitent et rendent compte du choix opéré. Les auteurs rédigeant des autobiographies « classiques » n'éprouvent pas le besoin de légitimer leur acte. Dans la situation cognitive particulière de ce séminaire, ces deux modèles de référence (non repérés et formulés comme tels d'ailleurs par les parties prenantes) sont disponibles, et finalement présents dans les pratiques réalisantes des auteurs.

C’est en cet endroit de l'analyse que nous pouvons nous appuyer sur le concept d'épreuve. Dans une telle perspective, le genre n’apparaît pas uniquement comme une codification plus ou moins contraignante à l'œuvre dans une culture située, et accessible à l'impétrant, mais également comme la résultante de l'expérience effective dans un cours d’action déterminé. La relative stabilité des caractéristiques du genre serait l'effet en retour de la constance de l'expérience effectuée. Les textes rassemblés dans notre corpus montreraient ainsi deux façons différentes de « traverser » l’épreuve, l’une s’appuyant sur la réception antérieure des récits autobiographiques, et s’absorbant dans l’activité du récit, l’autre opérant une transaction entre ce modèle reçu et celui du texte universitaire, ce dernier devenant alors genre d’accueil et imposant ses normes propres au matériel biographique, en rompant dans certains cas jusqu’à la chronologie et la mise en intrigue que le récit permet ou requiert. Les régularités observées dans notre corpus seraient en conséquence la résultante de la constance de l’épreuve, à l’évidence assurée par la stabilité du contexte. Si nous gardons présent à l’esprit que rédiger un texte autobiographique constitue sans doute pour la plupart des participants une activité nouvelle, nous pouvons identifier alors les composantes de ces deux façons de traverser l’épreuve, l’une reposant sur une réception large et plausible de récits autobiographiques, favorisant une identification des lois du genre et s’appuyant sur une compétence narrative antérieure, l’autre ne méconnaissant évidemment pas cette culture, mais « rapatriant » une compétence analytique dans le champ de l’autobiographie, compétence forgée pour une part dans ce même cadre universitaire.
Au plan du rapport entre structure des textes et contraintes des genres, il n’y aurait pas deux logiques distinctes, l’une diachronique, c’est-à-dire historique et représentée dans l’intertexte disponible, s’appuyant sur un héritage culturel correspondant aux formes actuelles des normes autobiographiques (récit d’enfance, galerie de portraits familiaux, etc.) et l’autre synchronique, développant une compétence analytique permettant de « traverser l’épreuve ». Dans notre perspective, un « énoncé concret » est toujours la résultante d’une épreuve et s’appuie à la fois sur des dimensions diachroniques, c’est-à-dire le recours sur des formes héritées, ici le récit autobiographique ou l’écrit universitaire, et synchroniques, c’est-à-dire étant perçues comme adaptées voire requises par la nature particulière de l’épreuve, c’est-à-dire de l’expérience effectuée.
Nous insistons sur cet aspect de l’épreuve qui nous paraît montrer que l’on peut ranger légitimement dans une catégorie générique identique (ici l’autoportrait) des textes produits en des époques ou des lieux différents, sans que l’on puisse supposer une filiation directe ou indirecte authentifiée. Ce qui permet de rapprocher l’autoportrait de la catégorie du *speculum*, c’est l’identité structurelle de l’épreuve, consistant ici à se penser ou se décrire selon les arcanes des lieux communs propres à une culture évidemment toujours située. L’épreuve n’est pas sans lien avec le contexte, mais elle ne s’y réduit pas et permet au contraire de rapprocher des situations institutionnelles différentes. Dans le cas du récit autobiographique, les auteurs de notre corpus ne traversent pas une épreuve différente que celle de tout autobiographe : il y a toujours une temporalité à organiser, une galerie de personnages familiaux à esquisser, un récit d’enfance à entreprendre, etc. Le contexte universitaire ne change rien à ces dimensions. L’hypothèse d’un lien étroit entre stabilité des genres et proximité d’épreuves rapprochées par leur structure commune conduit à envisager, dans le processus de sélection d’un modèle textuel par le sujet scripteur, l’influence de deux dimensions, celle bien connue de l’intertexte, où divers prototypes génériques sont identifiés et disponibles pour l’agent, et celle d’un inter-contexte, où des situations différentes sont rapprochées en fonction de certaines homologies pratiques.

5. Formes de mimèsis et économie cinétique

Une des caractéristiques du récit autobiographique est de reposer sur un pacte référentiel, qui a été décrit par Lejeune comme le fondement même du genre. Notre intention ici n’est pas de décrire les formes que peut adopter cet engagement, mais plutôt d’en examiner les conséquences au plan de l’organisation des temporalités, lesquelles subsument les formes de mimèsis qui en résultent. En effet, on a vu que l’un des gestes primordiaux de toute autobiographie est d’introduire un ordre dans le flux d’une vie, en périodisant celle-ci, sur la base de permanences recherchées. Le repérage de ces périodisations ne pose pas de difficultés majeures, dans la mesure où les textes livrent toujours des éléments de codage temporel, lesquels sont souvent extrêmement fins. Qu’on en juge par les exemples suivants :

- Je vois le jour, deux ans après la fin de la guerre, le 29 décembre 1947
- J'ai 15 ans
- L'année de ma matu [équivalent romand du "Bac" français], en 66
- J'accepte ce nouveau défi en 1990, avec plaisir ; etc.

Dans le présent article, nous souhaitons surtout privilégier les contraintes que les périodisations font peser sur les représentations de l'action. Ce travail suppose de s'appuyer sur un instrument narratologique pour analyser de près le fonctionnement de ces contraintes : il s'agit de la notion de *vitesse* du récit.

> On entend par vitesse le rapport entre une mesure temporelle et une mesure spatiale (tant de mètres à la seconde, tant de secondes par mètres) : la vitesse du récit se définira par le rapport entre une durée, celle de l'histoire, mesurée en secondes, minutes, heures, jours, mois et années, et une longueur : celle du texte, mesurée en lignes et en pages (Genette, 1972, p. 123).

L'analyse des « variations » de vitesse d'un récit (son économie cinétique donc) permet ainsi d'appréhender le traitement différentiel des périodes biographiques, le fonctionnement de la mise en intrigue propre au récit et les contraintes pesant sur les représentations de l'action. La numérisation du corpus étudié permet de recourir aux systèmes de comptage des traitements de texte et des tableurs, en disposant d'une évaluation fine du nombre de caractères. Le codage cinétique d'un texte ne pose donc pas de difficulté particulière. Il permet d'appréhender en première approximation le régime cinétique de chacun des textes, c'est-à-dire le rapport entre le temps chronique propre à la visée référentielle (l'étendue d'une vie, soit très simplement l'âge de l'auteur) et le nombre de caractère utilisé. Si par convention, on fixe comme base la page à 2500 caractères, nous avons le tableau suivant (les autoportraits ne sont ici pas pris en compte) :

Texte et auteur	Temps chronique représenté en moyenne par page
4 textes	± 18 mois
3 textes	± 2 ans
5 textes	± 3 ans
2 textes	de 4 à 5 ans
2 textes	6 ans
1 texte	12 ans

Tabl. 1 : Repérage des régimes cinétiques

6. Les contraintes narratives de l'économie cinétique

L'intérêt premier d'une telle représentation est de rendre accessible la diversité de l'économie cinétique d'un texte à l'autre, puisque celle-ci en moyenne peut varier de 1 à 8. Le codage des périodisations permet d'observer cependant que la vitesse d'un texte n'est jamais constante, mais au contraire connaît aussi une grande variété interne.

Quelles sont les formes repérées par la narratologie pouvant rendre compte des variations cinétiques ? Les travaux de Genette déjà cités en repèrent quatre : la pause, la scène, le sommaire et l'ellipse. La pause correspond à une suspension de l'action et accueille le plus souvent une description plus ou moins développée (le trait descriptif en constitue le point de repère spécifique) et/ou un commentaire du narrateur. La scène réalise conventionnellement (*ibid.,* p. 129) « l'égalité de temps entre récit et histoire » et peut accueillir des dialogues. On peut faire l'hypothèse que cette forme de mimèsis provient du genre dramatique et de l'action théâtrale, où l'on suit pas à pas l'action des personnages. Le « sommaire » est un récit résumé et présente une grande souplesse, dans le mesure où il couvre tout le champ compris entre la scène et l'ellipse. Il est une traduction du « *summary* » de la critique anglaise. Il assure le plus souvent la transition entre deux scènes, et constitue aux yeux de Genette le « tissu conjonctif par excellence du récit romanesque » (*ibid.,* p. 131). L'ellipse correspond comme chacun sait à un temps de l'histoire élidé, implicitement ou explicitement.

Le codage des périodisations permet de schématiser les variations de vitesse au fur et à mesure du développement des textes. Le tableau suivant propose une représentation de ces variations cinétiques pour une sélection de sept textes, retenus précisément pour leur grande diversité..

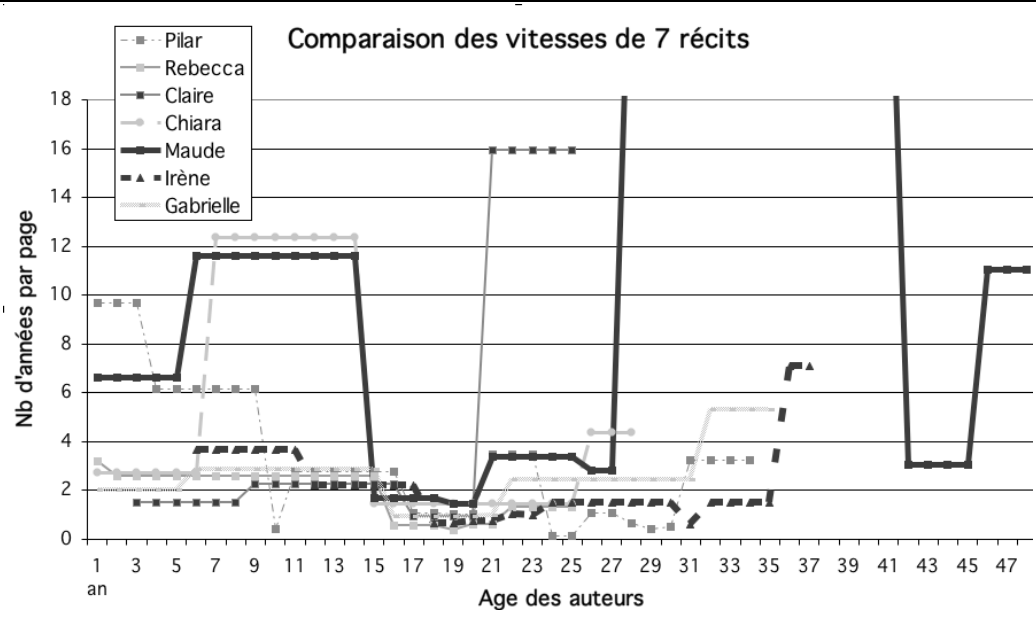


Fig. 1 : une schématisation de l'économie cinétique des récits

Deux points peuvent être soulignés à l'examen de ce schéma. Tout d'abord, il permet d'appréhender la diversité interne à chaque récit : tous les récits connaissent des changements de tempo, lesquels peuvent être abordés en première analyse comme les linéaments de la mise en intrigue propre à chaque récit : dès que l'on se rapproche de l'ordonnée, c'est-à-dire des valeurs lentes, on a affaire aux épisodes traités avec soin et détail, comme ici par exemple les 10 ans de Pilar (voir extrait 4) ou les 31 ans d'Irène. Inversement, on peut observer les valeurs extrêmement rapides de la période des 19/25 ans de Claire, ou des 27/41 ans de Maude et faire l'hypothèse inverse : les choix narratifs opérés éloignent ces périodes de la mise en intrigue construite par le récit. Observons au passage que *tous* les récits schématisés adoptent des valeurs lentes pour l'intervalle situé entre 15 et 24 ans. Il y a certes toujours des singularités, c'est-à-dire que des périodes situées hors de cet intervalle feront l'objet d'un traitement approfondi, mais tous les textes comporteront des narrations développées d'épisodes se situant dans la jeunesse des auteurs, au point que celle-ci puisse apparaître dans notre corpus comme l'apogée du récit autobiographique : le récit de vie est d'abord un récit de la jeunesse. Pour le dire vite, l'intrigue de l'adulte serait une herméneutique de sa jeunesse : tel serait le lieu commun de l'autobiographie et la formation culturelle partagée par nos auteurs.

7. Récit sommaire et syllepses

Mais l'intérêt premier de ces schématisations ne nous semble pas résider en ces propositions sans doute trop spéculatives, car il permet surtout d'observer de près les compétences narratives déployées et les formes de mimèsis qui en résultent. Pour ce faire, nous avons besoin d'un ultime instrument :

> On pourrait nommer *syllepses* (fait de prendre ensemble) *temporelles* ces groupements anachroniques commandés par telle ou telle parenté, spatiale, thématique ou autre (Genette, *op. cit.*, p. 121).

Par groupements « anachroniques », il faut bien entendu comprendre des rapprochements d'actions ou d'événements itératifs indépendamment de leur calendrier de mise en œuvre ou d'apparition : la scolarité, les loisirs, la famille, les voyages, etc. Elles peuvent être aussi topologiques ou temporelles : ce qui se passait dans tel lieu, ou à telle époque. Si l'on accepte la définition nouvelle du terme de syllepse[1], dans cette acception de « prendre ensemble », alors nous disposons d'un instrument d'analyse qui rend compte de la grande variation cinétique de nos récits. L'extrait 5 montre un exemple de valeur cinétique et permet de définir le fonctionnement du summury « rapide » dans notre corpus : une seule syllepse, celle de la scolarité, couvrant une

[1] En rhétorique, ce terme désigne un trope particulier. Soulignons donc, pour éviter toute confusion, que l'acception de la syllepse sera ici narratologique et non pas rhétorique.

longue période de près de quinze année, expédiée en quelques lignes. La compétence narrative usuelle telle qu'elle est repérable dans notre corpus recourt massivement au récit sommaire. Lorsqu'une période fait l'objet d'un traitement soigné et comporte en conséquence des valeurs extrêmement lentes, on assiste à une multiplication des syllepses portant sur des séquences temporelles beaucoup plus courtes. C'est ce jeu simultané du nombre important des syllepses et de l'empan temporel retenu qui rend compte de la variabilité cinétique. L'extrait 6 présente le fragment d'un récit sommaire de type lent, relatant un séjour de près de 20 mois réalisé à l'étranger, alors que l'auteur a 16 ans, et qui fera l'objet d'un récit regroupant une douzaine de syllepses, dont la plupart permettent de relier des dimensions itératives : l'accord du père et de la mère, l'adaptation socio-culturelle, le lycée binational, la réussite du Bac, l'initiation socio-politique, les voyages dans le pays, les copains, etc. : l'amplification d'une période repose sur la multiplication des syllepses.

8. Récit sommaire et mise en intrigue dans le corpus

Les formes de mimèsis adoptées par nos auteurs conduisent à un fonctionnement particulier de la mise en intrigue. Dans la tradition romanesque, il est indubitable que celle-ci opère électivement sur des *scènes*, qui fournissent la matière primordiale du récit, et non pas sur les formats de récit sommaire, qui assurent les transitions nécessaires entre les scènes, et qui se caractérisent au plan de l'isochronie entre temps du récit et temps de l'histoire par un découplage total (c'est-à-dire que plusieurs heures, journées, semaines, etc., sont « traitées » en quelques lignes). La difficulté de notre corpus tient au fait que le recours à la scène y est bien attesté (extrait 4) mais rare, alors que l'usage du *summury* y est quasi-permanent.

Le critère de vitesse est ici précieux, car il fournit des indications sensibles permettant de repérer les « apogées du récit », c'est-à-dire les actions ou événements qui font l'objet d'un traitement soigné, soit par le recours au format scénographique, soit (c'est le cas le plus fréquent) par la mise en œuvre d'un *summury* à syllepses multiples. Notre hypothèse est la suivante : il n'est pas envisageable que des actions et événements faisant l'objet d'un développement approfondi ne jouent pas un rôle majeur dans le dispositif de mise en intrigue. Ils le *signalent*, orientent la lecture et les directions de l'interprétation. Amputer le récit des sections ainsi repérées, c'est sans doute plus que le mutiler, c'est le trouer de telle sorte que des logiques d'action échappent à la compréhension.

La détermination des vitesses permet de repérer grossièrement les « épisodes » qui importent dans le dispositif de mise en intrigue.

Les syllepses fournissent un indicateur aisément repérable de l'organisation de l'activité narrative. Elles sont d'une grande diversité. Certaines sont à l'évidence communes et ordinaires : syllepse du scolaire, syllepse de la vie de famille dans les périodes de l'enfance et de l'adolescence, syllepse des vacances, syllepse des voyages, syllepse des « premiers souvenirs d'enfance ». Il y a des syllepses curieuses, nullement « obligées », et qui pourtant, par leur émergence dans pratiquement un récit sur deux, finissent par prendre aux yeux du lecteur ce caractère « obligé » : ainsi de la syllepse d'une figure professorale privilégiée, tel par exemple le « maître de mes 9 ans », dans un des textes. Mais aux côtés de telles « figures imposées », si l'on peut dire, les syllepses ordinaires correspondent sans doute aux « encyclopédies de savoirs partagés », c'est-à-dire les connaissances nécessairement communes aux locuteurs pour entrer dans la compréhension élémentaire d'un énoncé. Certaines font époque : comment penser la syllepse des vacances avant l'invention des congés payés ? D'autres semblent d'une longévité d'un autre ordre, comme celle des amours.

Dans tous les cas, les syllepses retenues par les auteurs nous informent au plan culturel, et l'on saisit mieux alors une dimension épistémique des autobiographies, qui tient précisément à leur valeur de témoignage culturel, et dont les « syllepses » nous permettent une sorte de repérage basique et peut-être primordial. Il y a dans certains textes par exemple des syllepses du monde de « l'avant », propre à celui des années 40 et 50 : la récupération du « révolu » est une dimension fondatrice du courant autobiographique.

Mais là encore, tout comme pour notre acmé du jeune adulte, des *singularités* ne sont pas exclues, évidemment *relatives*, mais vraiment propres à un auteur. Ainsi de la syllepse de la souffrance dans un des récits, unique dans notre corpus, mais suffisamment commune pour être partagée. Ainsi de la syllepse des errances nocturnes d'un autre. Même remarque. Ainsi de la syllepse d'un âge aussi particulier que les « 10 ans », ou de celle de l'émigration (davantage

partagée évidemment) qui fait que l'on pourrait « mettre en série » plusieurs récits. Syllepse de la contradiction entre conformité et révolte pour un texte, qui peut faire écho pour d'autres. Syllepse des lieux investis encore. Syllepse du « ce n'est pas toujours mieux ailleurs ». Syllepse mystérieuse et comme souterraine (c'est-à-dire non explicite, et peut-être « involontaire ») de *la prime initialité* (privilégier dans son récit ce qui advient pour la première fois : la première prise de fonction, le premier mandat politique, le premier grand voyage, etc.).

9. Conclusion : de l'épreuve aux constantes génériques

La scène est rarement sollicitée comme formule narrative dans les textes de notre corpus. Dès qu'un auteur en use, il y a nécessairement infléchissement du régime cinétique vers des valeurs lentes et amplification ou dilatation textuelle corrélative. On peut comprendre la « raison » de cet usage parcimonieux en prenant en compte le fait qu'une utilisation massive du format scénographique générerait des textes volumineux, demandant un temps de rédaction incompatible avec celui à disposition. Nous retrouvons ici la notion d'*épreuve* vue plus haut, où la stabilité des traits caractéristiques d'un genre résulte *également* de l'identité de l'expérience effectuée. Certaines contingences peuvent entrer dans la structure de l'épreuve, comme à l'évidence le temps disponible pour la rédaction d'un texte. Nous pouvons supputer avec quelques vraisemblances qu'une transaction est opérée par chaque auteur entre la loi du genre et cette contingence, c'est-à-dire entre la visée de la totalité biographique à couvrir et les heures disponibles pour honorer cette visée. Cette situation particulière conduit ainsi à l'adoption massive du *summury*, dont la plasticité diachronique permet de traiter un empan biographique de grande envergure. Sur ce plan, un texte du corpus comporte une curiosité, puisque l'auteur a cru bon de laisser figurer, en fin de texte, une indication informatique comprenant nom de fichier, nombre de minutes et de mots. Nous apprenons ainsi que l'auteur a passé quelques 29 heures à rédiger son texte. Indication certes superficielle, mais révélatrice d'une contingence contraignante. Les formes de mimèsis ne s'émancipent donc pas totalement des caractéristiques de l'activité réelle qui les produisent, et l'on peut tenir pour plausible l'hypothèse d'une loi nécessaire reliant temps à disposition, ampleur des formats produits et formes de mimèsis adoptées. Nous aurions ainsi deux formes de l'épreuve autobiographique : (i) je dispose de beaucoup de temps, je produis un texte ample (≥ 100 pages), je recours massivement au traitement scénographique ; (ii) je dispose de peu de temps, je ne peux (ou ne dois) que produire un texte au format modeste (≤ 20 pages), et je recours massivement au *summury*.

Le plan sémiotique se définit comme une structure « ouverte », dépendant des pratiques qui le mobilisent et le mettent au travail. Dès lors que l'hétéronomie de la chose langagière est admise, on est conduit à rechercher les formes de régulation de celle-ci. En premier lieu du côté du couplage texte/cours d'action. Dans le cas de l'autobiographie, notre concept d'*épreuve* vise à rendre compte des régularités du genre, dans une orientation hétéronomique, venant ainsi compléter l'hypothèse (trop forte à nos yeux) que la régularité des genres dépendrait uniquement des pratiques réceptives. Notre concept d'épreuve suppose l'existence d'une formation culturelle partagée, qui donne sens et plausibilité à la réalisation d'un genre. Dans le cas de l'autobiographie, cette formation culturelle conduit à envisager la vie d'une personne comme une « totalité » dont on peut faire le récit. Notre concept d'épreuve stipule simplement ceci : dès lors qu'un sujet se met dans la situation propre à la réalisation d'un genre, il est conduit à mettre en œuvre quelques gestes fondamentaux, correspondant à la structure de l'épreuve, et contribuant à la régularité des formes génériques. Dans le cas de l'autobiographie, Lejeune a remarquablement décrit ces gestes : recherche de permanences et introduction d'un ordre, afin selon nous de réduire le problème majeur de l'épreuve, qui est le risque de *prolifération de l'information biographique*. Il y a tant à dire qu'il est *nécessaire* d'adopter des principes régulateurs. Le concept d'*épreuve* nous semble correspondre très exactement à cette *nécessité*. La régularité des genres, dans une telle perspective, apparaît ainsi comme l'effet de deux facteurs : (i) la réception antérieure par le sujet de textes analogues, produisant une culture diffuse du genre considéré, que l'on peut conceptualiser en termes de « tradition », et permettant un jeu permanent de prorogation ou de renouvellement ; (ii) la distribution de situations identiques « provoquantes », dont la structure commune (rendre viable le fait de raconter sa vie, par l'introduction d'un principe d'ordre) favorise le maintien des « lois du genre ». *Nous abordons ainsi les performances textuelles à la confluence de deux ordre de détermination : la réception antérieure de textes en une culture*

donnée ; l'implication en des cours d'action distribuant des épreuves communes et dépendant pour une part des pratiques sociales qui les organisent.

BIBLIOGRAPHIE

BAKHTINE, M. 1984. *Esthétique de la création verbale*, Paris, Gallimard.

BAUDOUIN, J.-M. 2001. La dimension du groupe, seconde et primordiale : histoire de vie et recherche-formation, in C. Solar (éd.), *Le groupe en formation*, Bruxelles, De Boeck, pp. 35-56.

BEAUJOUR, M. 1980. *Miroirs d'encre. Rhétorique de l'autoportrait*, Paris, Seuil.

CHISS, J.-L. 1985. Raconter et témoigner : le vécu à la croisée du théorique et du politique, *Pratiques,* 45, pp. 13-31.

DEMAZIÈRE, D. & DUBAR, C. 1997. *Analyser les entretiens biographiques*, Paris, Nathan.

GENETTE, G. 1972. *Figures III*, Paris, Seuil.

GENETTE, G. 1983. *Nouveau discours du récit*, Paris, Seuil.

LAINÉ, A. 1998. *Faire de sa vie une histoire. Théories et pratiques de l'histoire de vie en formation*, Paris, Desclée de Brouwer.

LE GRAND, J.-L. & PINEAU, G. 1993. *Les histoires de vie*, Paris, PUF.

LEJEUNE, Ph. 1975. *Le Pacte autobiographique*, Paris, Seuil, Coll. Points.

LEJEUNE, Ph. 1996. L'ordre d'une vie, in M. Contat (éd.), *Pourquoi et comment Sartre a écrit " Les Mots "*, Paris, PUF, pp. 49-120.

PENEFF, J. 1990. *La méthode biographique*, Paris, Armand Colin.

POIRIER, J., CLAPIER-VALLADON, S. & RAYBAUT, P. 1983. *Les récits de vie. Théorie et pratique,* Paris, Puf.

RASTIER, F. 2001. *Arts et sciences du texte*, Paris, PUF.

ANNEXES

Extrait 1

En 1959, l'Espagne vivait sous la dictature de Franco. Les régions du sud étaient défavorisées par rapport à celles du nord où toute l'industrie s'était installée.

Luis et Pilar, après leur mariage et en guise de voyage de noces avaient émigré vers Balaguer un village de Lérida en Catalogne, où lui avait déjà trouvé du travail avant de partir pour se marier.

Pour Pilar ce village était presque l'étranger, bien qu'elle ait déjà habité à Madrid pendant trois ans, mais ici c'était encore différent: elle ne connaissait personne, elle ne comprenait pas la langue de gens du village. Avant d'être venue, elle ne savait même pas qu'il existait en Espagne des régions où leur population ne parlaient pas le castillan. Pour elle, comme pour la plupart des espagnoles pendant le franquisme, l'Espagne était une, seule, et grande.

En plus dans le village, elle n'avait rien à faire, à part de s'occuper de son mari, qui partait souvent en voyage pour cinq ou six jours. Pilar, ma mère, m'a à peine parlé de cette époque, mais elle m'avait fait comprendre qu'elle avait été une époque difficile.

Extrait 2

La figure du maître

Le thème central de mon récit est celui de la figure du maître. Nous apprenons de multiples façons, dans des livres et par l'expérience, à l'école et dans la souffrance, dans les voyages et dans l'activité professionnelle. J'ai fréquenté tous ces lieux de l'apprentissage, j'en ai investi certains plus que d'autres. Mais lorsque je pense aux mots "grandir", "apprendre", "formation de l'identité", etc, je vois d'abord des visages; ceux de personnes qui ont marqué différentes étapes de mon existence; des visages que j'ai reconnus dans une foule d'autres visages et auxquels je me suis identifié, auxquels j'ai voulu ressembler ou que j'ai voulu suivre. Pourquoi ces visages particuliers et pas d'autres ? La question reste ouverte devant l'insondable complexité des relations humaines et de la construction de l'identité psychique.

Extrait 3

Piaget avait décrit le développement cognitif de l'enfant, qui part d'un point de vue autocentré, pour se décentrer peu à peu. Il illustrait parfaitement ce processus par exemple avec l'expérience des trois montagnes. Piaget a décrit ce processus d'un point de vue cognitif. Mais je crois qu'il s'agit d'un phénomène de décentration beaucoup plus large. D'après Schmidt Kitsikis (UNIGE), cette possibilité de décentration cognitive nécessite un développement similaire au niveau affectif, qui va du narcissisme (amour de soi) vers la relation d'objet (amour de l'autre). Ce processus est censé être achevé à l'adolescence selon les théories classiques. Je ne partage pas ce point de vue. Car je sens très bien que pour moi, ce développement est toujours en cours. Il est même assez récent. La décentration signifie concrètement la capacité de prendre en compte un point de vue différent du sien, donc la capacité à relativiser son propre point de vue.

Extrait 4

Pour Noël nous allons passer les fêtes dans l'appartement que nous avons encore à Lérida. La nuit du 30 au 31 ma mère est arrivée au terme de sa grossesse et on attend l'accouchement d'un moment à l'autre. Ce nuit-là, nous dormons les deux seules dans l'appartement, j'ai la charge, si elle commence à avoir mal, d'avertir mon oncle qui habite à l'étage inférieur.

Il fait très froid dans l'appartement vide depuis des mois. J'ai mon nez froid, je ne peux pas dormir mais ce n'est pas à cause du froid, je n'arrête pas d'imaginer comme cela va arriver. Je l'entends se plaindre de temps en temps, je suis inquiète, elle ne me laisse pas avertir.

A 7h après avoir entendu sonner le réveil de mon oncle, elle me dit d'avertir. Mon oncle monte tout de suite et il appelle un taxi pendant qu'elle s'habille lentement en gémissant par moments.

Je suis resté collée à la vitre. La rue est toute verglacée, toutes les voitures et les maisons sont pleines de neige j'observe étonne ce spectacle peu fréquent dans notre ville.

Extrait 5

Ma mère s'occupant de nous chaque soir pour les devoirs, j'ai vite été à l'aise à l'école. En deuxième primaire, j'ai eu la possibilité de sauter une classe. Le revers de cette facilité fut que je n'ai jamais vraiment choisi mon parcours scolaire. Mes notes me guidant, j'ai ainsi effectué un cycle d'orientation en section latine et un collège en section classique (grec, latin).

Extrait 6

Mes débuts là bas [en Amérique du Sud] ont été bizarres. D'abord parce que la rapidité de ma décision ne m'avait pas donné le temps de réaliser pleinement ce qu'il m'arrivait, ensuite, parce que je débarquais dans un monde qui m'était totalement étranger. Mon seul point de repère était Zara, tout le reste était à découvrir. En commençant par la langue, la ville et ses alentours, les gens du pays, les normes sociales, les "que dire à qui, comment et pourquoi", et tout ce que doit savoir une personne désirante de s'adapter à son nouvel environnement.

Je me souviens de ma sensation "d'avoir reçu un coup de massue sur la tête" les premiers mois de mon séjour, tant il y avait d'informations et de nouveautés à emmagasiner. J'étais à la fois

émerveillée et déchirée par les réalités [du pays], et toujours à la recherche d'en savoir un peu plus. J'avais envie de me donner tous les moyens pour saisir ce qui m'était inconnu, étranger, incompris ou insupportable. Aujourd'hui, je n'ai toujours pas tout compris, mais je pense que les moyens que je me suis donnés pour chercher à comprendre m'ont permis de découvrir de vivre et d'apprendre beaucoup d'autres choses. (= petite minute de philo pas chère).
Mais avant de m'investir plus à fond dans mon adaptation socioculturelle, il me fallait porter un oeil sur les études et réussir le bac en quelques mois. La R… était le lycée bi¬national de G…, connu pour ses hautes exigences scolaires et dont le papier de sortie était l'un des mieux coté du pays. Il s'agissait en fait d'un lieu à moitié burlesque pour une suissesse: les classes ressemblaient à des bungalows, les intérieurs étaient enfumés par les profs et les élèves qui se passaient des clopes tout zazimut, et un lama fétiche broutait non stop l'herbe du préau. Il m'était difficile de ne pas réaliser que j'étais dans les Andes! Sans oublier les interruptions des cours à chaque passage d'un avion: G… avait eu la bonne idée de construire son aéroport au centre ville.

DIACHRONIE COMPARÉE DE FORMES MÉSO ET MACRO-SÉMANTIQUES DANS LE CORPUS *GILLES DELEUZE*

Sylvain LOISEAU
MoDyCo, Paris X

SOMMAIRE

Résumé : *Alors que les corpus richement annotés renouvellent la description sémantique, on souhaite dans cette contribution utiliser ce nouveau type d'objet empirique pour décrire la diachronie de formes sémantiques. Deux types de formes sémantiques en particulier sollicitent notre attention : le thème sémantique d'une part, le genre d'autre part. On essayera de décrire les traits caractéristiques de leur évolution, et l'articulation de leur évolution. Le corpus étudié comprend l'intégralité des commentaires philosophiques et des essais de Gilles Deleuze. L'empan diachronique s'étend de 1953 à 1994, c'est-à-dire de la philosophie d'après-guerre jusqu'à l'apparition de la génération philosophique d'après 1968. On s'attachera d'abord à justifier une périodisation qui distingue trois pôles : l'académisme des années 50, le militantisme des années 60 et 70, et le retour à un certain classicisme à partir des années 80. De nombreux traits contribuent à justifier cette périodisation, aussi bien dans les systèmes thématiques et dialogiques que, à un autre niveau, à travers des traits morphosyntaxiques et stylistiques.*
À partir d'une caractérisation des textes relevant de chacun des deux genres, le commentaire et l'essai, sur cet axe diachronique, la comparaison de la diachronie des genres permet d'observer des évolutions corrélées. La période centrale est particulière du fait de la disparition apparente du genre du commentaire de la production de Deleuze. On montrera comment les deux genres ont temporairement été confondus, dans un projet de renouvellement et de sommation des formes textuelles. Enfin, on s'attachera à montrer la corrélation d'une forme théorique caractéristique de la période du début des années soixante aux formes textuelles dans lesquelles elle s'inscrit. Le concept de différence est caractéristique d'une remise en cause de l'ontologie par la philosophie de l'après guerre d'Algérie, commun notamment à Deleuze (Différence et répétition, 1969), Derrida (L'Ecriture et la différence, 1979), et Lyotard (Le Différend, 1983). Sur le sous-corpus constitué de ces trois textes on exploitera les corrélations entre les différents systèmes linguistiques annotés pour montrer l'apparition et la construction d'une thématique partagée, centrale pour l'histoire des idées.

Introduction

Les corpus annotés peuvent-ils aider l'analyse diachronique ? Dans le cadre d'une recherche de méthodes, on souhaiterait faire porter cette question sur un point précis : celui de l'articulation de description diachronique de formes sémantiques à des paliers d'analyse différents[1].

Le corpus étudié comprend l'intégralité des commentaires philosophiques et des essais de Gilles Deleuze. L'empan diachronique s'étend de 1953 à 1994, c'est-à-dire de la philosophie d'après-guerre à l'apparition de la génération philosophique d'après 1968 (les références des textes du corpus sont données en annexe).

[1] Je remercie Céline Poudat pour son aide.

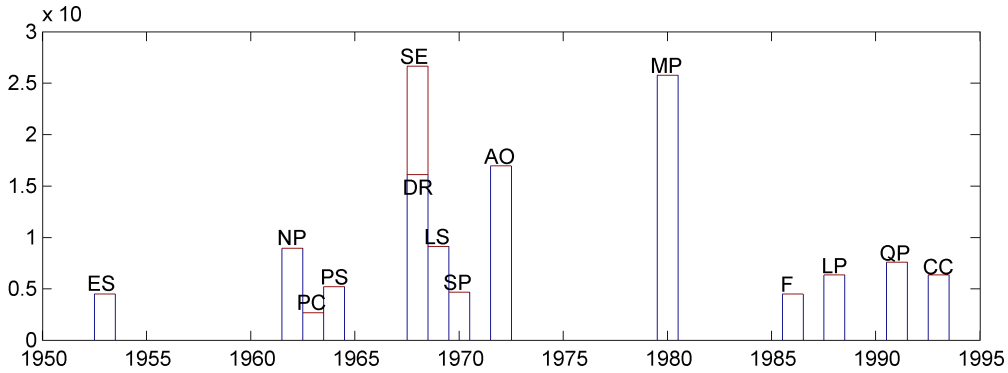


Figure 1 : distribution chronologique et taille (en nombre de mots) des textes du corpus

On se limitera dans les résultats présentés ici à l'utilisation de variables morphosyntaxiques pour se concentrer sur l'observation de l'évolution du système dialogique des textes du corpus. L'observation portera sur deux paliers différents : celui des textes et celui des paragraphes, afin d'essayer de décrire l'évolution de la structure globale du corpus et de la structure interne des textes.

1. Variables utilisées

Dans cette analyse on s'attache à l'observation des textes du corpus sur des critères essentiellement morphosyntaxiques. On a privilégié des variables reflétant le système verbal (temps et personne). Ont été ajoutés les ponctuations, les types de noms (singuliers, pluriels, ou propres), et l'opposition fini/indéfini (relevé sur les pronoms et les déterminants). Enfin on a également inclu les effectifs des notes, citation, mises en forme, titres et segmentations des textes en sections et sous-sections. De nombreux tests ont permis de montrer que l'hétérogénéité de ce jeu de variables ne fait pas sensiblement baisser la qualité de leur regroupement en facteurs : les deux premiers facteurs résument 51% de la variance quand sont utilisés seulement les temps et les personnes, tandis qu'ils résument 52% de la variance sur l'ensemble de ce jeu de descripteurs.
Les variables ont été constituées d'une part à partir d'un balisage des propriétés structurelles du corpus et d'autre part à partir d'une annotation morphosyntaxique du logiciel Cordial, importée dans l'annotation existante[1]. Les étiquettes de Cordial ont été décomposées, dans le corpus, en traits minimaux (partie du discours, genre, nombre, caractère fini ou indéfini d'un pronom ou d'un déterminant, etc.). Les descripteurs utilisés ont été constitués « à façon », par un assemblage de ces traits. Par exemple, on a utilisé comme descripteur « noms communs singuliers », par un assemblage de plusieurs traits issus de l'analyse de Cordial (« Nom commun singulier » est la cooccurrence du trait « nom » et du trait « singulier » sur un mot).
Ce jeu de variables est sémiotiquement hétérogène : il relève à la fois de la syntaxe (à travers les temps et les personnes), de propriétés éditoriales ou macrostructurelles (segmentation en paragraphes, notes, etc.), et d'un niveau peut-être « rhétorique », à travers les ponctuations. Les paliers observés sont également très hétérogènes : les segmentations ou les notes relèvent du macrosémantique, tandis que l'opposition entre défini et indéfini par exemple est observée au palier des mots. Les ponctuations relèvent peut-être davantage du palier mésosémantique (palier de la période)[2].
Enfin, ce type de descripteurs pose des difficultés particulières. Les métriques sont très différentes : la moyenne la plus élevée est de 73,63 (nom singulier), tandis que la moyenne la plus faible est de 0,13 (point d'exclamation). Mais surtout, ces descripteurs forment des sous-ensembles (temps, personnes), et il est intéressant de pouvoir prendre en compte le caractère structurel de ces sous-ensembles, plutôt que de traiter uniformément les fréquences. À cette fin, on a utilisé la méthodologie établie par Céline Poudat (Poudat, 2006), qui consiste à soumettre à

[1] Cette importation, ainsi que l'extraction des fréquences des traits, a été réalisée avec le logiciel CorpusReader (Loiseau, 2006)
[2] Les descripteurs sont préfixés, dans cet article, par des accolades qui indiquent grossièrement leur origine sémiotique : {gram} indique un descripteur grammatical, {lm} un lemme, {ff} une forme fléchie, {c} un caractère de ponctuation, {tag} un élément de structuration. Ces préfixes ne prétendent cependant pas établir une typologie du système sémiotique ; ils reflètent originellement l'organisation matérielle de l'annotation du corpus.

l'analyse factorielle non pas des fréquences (absolues ou normées), mais des pourcentages dans différentes catégories comprenant les variables en distribution complémentaire. Ainsi, la valeur de la variable « imparfait » dans chacun des textes est le pourcentage des imparfaits dans l'ensemble des temps.
Pour les traits correspondant à des marques de segmentation (note, citation, titre, division, mise en forme), on n'a cependant pas pu opérer de regroupement en classes, ces variables n'étant pas en distribution complémentaire. On a alors utilisé un pourcentage calculé par rapport à la longueur des textes (mesurée en nombre de mots). L'interprétation des variables en dépend : la variable « nombre de notes » indique la « densité » en notes, tandis que la variable « nombre de paragraphes » indique davantage la longueur moyenne des paragraphes, puisque les textes sont entièrement segmentés en paragraphes. L'hétérogénéité du jeu d'étiquettes se répercute sur l'interprétation qu'elles permettent.
En résumé, le jeu d'étiquettes utilisé se répartit comme suit :

Catégorie	Etiquette	Nom
type de noms	{gram}ty.p	Nom propre
	{gram}nom-s	Nom commun singulier
	{gram}nom-p	Nom commun pluriel
ponctuations	{c}?	Point d'interrogation
	{c}!	Point d'exclamation
	{c}.	Point
	{c},	Virgule
	{c}...	Points de suspension
	{c}:	Deux-point
	{c};	Point-virgule
Verbes	{gram}t.p	Présent
	{gram}t.i	Imparfait
	{gram}t.pas	Passé simple
	{gram}t.fut	Futur
	{gram}t.subpre	Subjonctif présent
	{gram}t.subimp	Subjonctif imparfait
	{gram}t.con	Conditionnel
	{gram}t.imp	Imparfait
	{gram}t.partpas	Participe passé
	{gram}t.partpres	Participe présent
Personne en lemme	{lm}je	Lemme *je*
	{lm}tu	Lemme *tu*
	{ff}il	Forme fléchie *il*
	{lm}nous	Lemme *nous*
	{lm}vous	Lemme *vous*
	{ff}ils	Forme fléchie *ils*
	{lm}on	Lemme *on*
fini ou indéfini	{gram}fin.d	Traits défini (sur un pronom ou un déterminant)
	{gram}fin.i	Traits indéfini (sur un pronom ou un déterminant)

Tableau 1 : Descripteurs regroupés en classes

{tag}note	Nombres de note
{tag}hi	Nombre de mise en forme (gras, italique, …)
{tag}head	Nombre de titre
{tag}div	Nombre de divisions et subdivisions
{tag}q	Nombre de citations
{tag}p	Nombre de paragraphes

Tableau 2 : Descripteurs hors classe

2. Analyse factorielle au palier des textes

Les variables ainsi calculées[1] sont soumises à une analyse en composante principale dans le logiciel DTM[2]. Les valeurs propres des premiers axes représentent un pourcentage très satisfaisant de la variance totale. La stabilité et l'individualisation des axes issus des descripteurs

[1] La manipulation des variables s'est effectuée à l'aide du logiciel Matlab.
[2] (Lebart, 2006)

sélectionnés sont confirmées par la longueur et le faible recouvrement des intervalles d'Anderson correspondants.

nbr	val. prp.	%	cumul. %	
1	10.8796	32.00	32.00	**********
2	6.7282	19.79	51.79	**********
3	4.1454	12.19	63.98	**********
4	2.9906	8.80	72.78	**********
5	2.4524	7.21	79.99	**********

Figure 2 : valeurs propres des cinq premiers axes

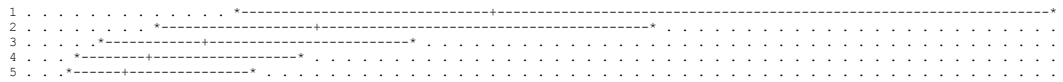


Figure 3 : représentation des intervalles d'Anderson des cinq premiers axes

La distribution des textes sur le premier facteur restitue notablement l'axe diachronique du corpus : l'origine de l'axe oppose les textes sur la coupure des années 1970. Les exceptions apparentes à cette distribution chronologique sont *Proust et les signes* (PS) d'une part, qui se positionne du côté des textes postérieurs, et *Spinoza et le problème de l'expression* (SE) d'autre part, qui se positionne avant *Différence et répétition* (DR). Dans le premier cas, on peut rapprocher le caractère irrégulier de la position de *Proust et les signes* (PS) de l'originalité de son objet dans le corpus. Dans le second, le recul sur le premier axe de *Spinoza et le problème de l'expression* (SE), est en partie un artefact de la numérotation des textes, puisque ce livre est paru la même année que *Différence et répétition* (DR) (et est issu, comme lui, d'une thèse de troisième cycle). La différence cependant sensible de leur position peut sans doute être rapportée à l'opposition entre genre du commentaire et genre de l'essai : *Différence et répétition* formule le programme qui sera réalisé dans les essais postérieurs. Les quatre derniers textes sont remarquablement regroupés. On peut noter, pour terminer, que l'opposition entre essais et commentaires est peut-être « en germe » dans le second axe : l'opposition entre les deux genres est contrariée seulement par la position de *Nietzsche et la philosophie d'une part* (NP), et celle de *Qu'est-ce que la philosophie* (QP) d'autre part.

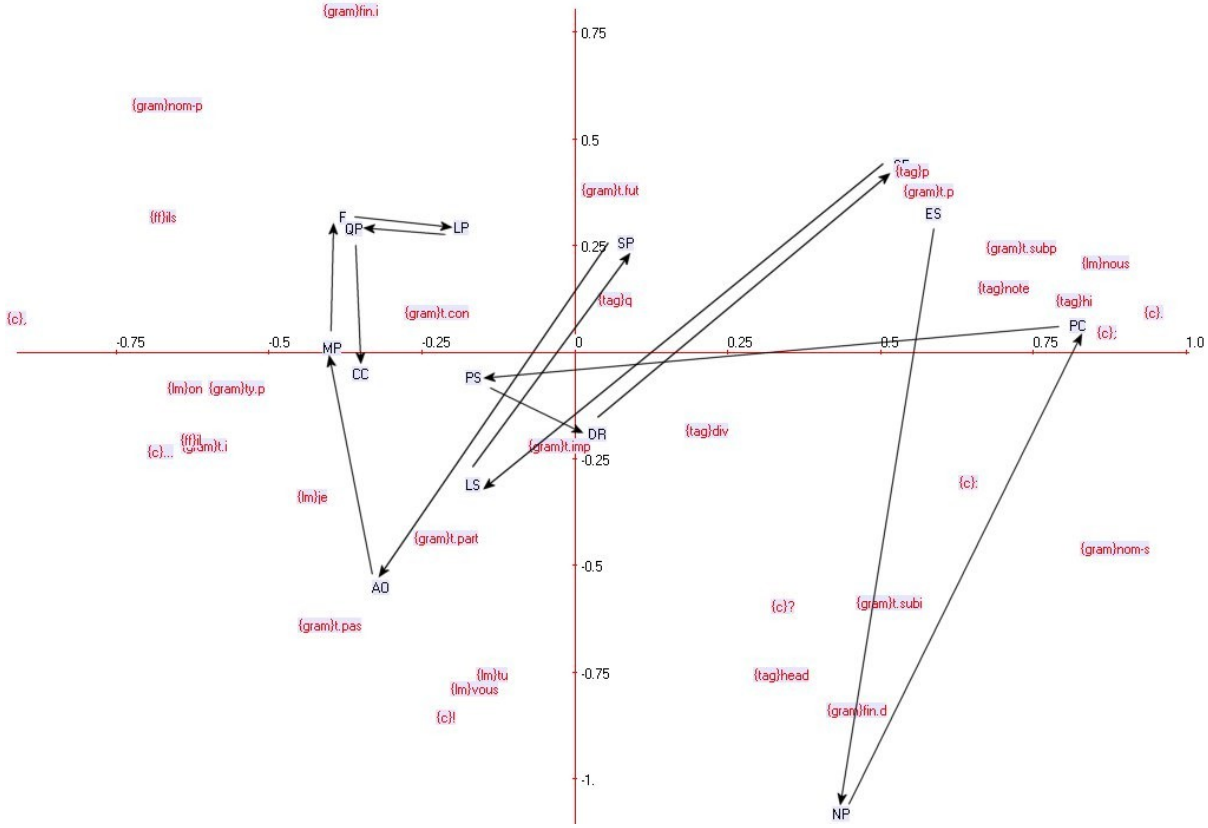


Figure 4 : ACP des textes du corpus. Les textes sont reliés dans l'ordre chronologique

Le premier axe semble organiser les variables autour d'une opposition entre un plus ou moins grand académisme : les marques de structuration des textes, les deux temps du subjonctif présent et passé, le *nous*, les ponctuations complexes comme les deux points et le point virgule, s'opposent notamment aux personnes *je*, *tu* et *vous*, aux points de suspension et d'exclamation.
L'arbre de classification[1] donne une représentation plus fine de la proximité entre textes. On observe que la plus grande distance est entre les quatre premiers commentaires du corpus et les textes suivants. Si l'on coupe cet arbre pour obtenir quatre classes[2], les regroupements constitués respectent donc l'axe chronologique : seuls *Spinoza et le problème de l'expression* (SE) et *Proust et les signes* (PS) perturbent cette répartition, puisqu'ils sont inclus dans une classe alors qu'ils sont situés chronologiquement entre deux textes d'une autre classe. Les perturbations de l'axe chronologique semblent donc relever d'une opposition entre genres.

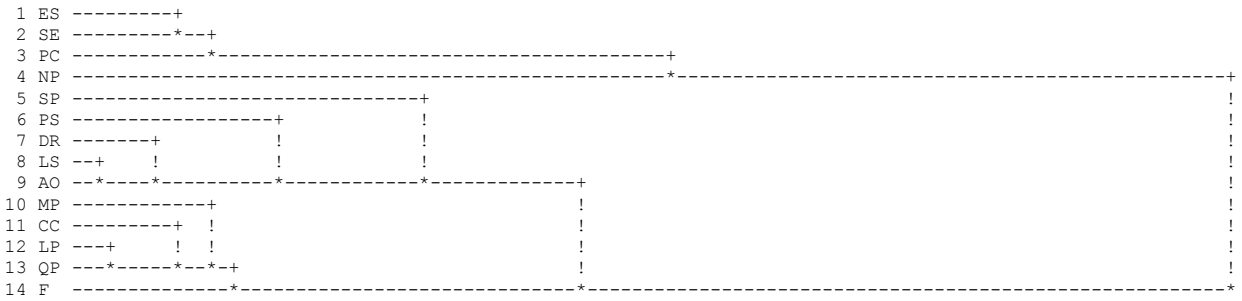


Figure 5 : Dendrogramme des textes

La classification des textes en quatre classes, issue de ce dendrogramme, permet d'observer, plus finement que sur les deux premiers axes factoriels, les variables spécifiques de chaque groupe.

Classe	Individus-textes	Variables positives[3]	Variables négatives
Classe 1	ES, PC, SE	points, points virgules, notes, mises en formes, *nous* et subjonctif présent.	virgule, points de suspension, *il*, *on*, noms propres, et passé simple.
Classe 2	NP	Pas de variables caractéristiques.	Pas de variables caractéristiques.
Classe 3	PS, DR, LS, SP, AO	Participe passé, passé simple.	Futur.
Classe 4	MP, F, LP, QP, CC	*On*, indéfini, noms propres, virgule, conditionnel, points de suspension.	Subjonctif imparfait, noms communs singuliers, défini, nous, points-virgules.

Tableau 3 : les quatre classes issues d'une classification hiérarchique et leurs variables caractéristiques

Restituant en grande partie l'ordre chronologique, l'analyse factorielle permet d'une part de proposer une segmentation diachronique du corpus, et d'autre part de caractériser chaque sous-ensemble.
Les bornes temporelles des trois classes sans singleton sont 1953 et 1968 (15 ans) pour le premier, 1964 et 1972 (8) pour le second, et 1980 et 1994 pour le quatrième (14). Ces bornes sont en recouvrement partiel. Elles sont hétérogènes en genre, à l'exception de la première.
Les oppositions des descripteurs se concentrent principalement entre les classes 1 et 4 – les classes les plus éloignées chronologiquement – qui regroupent respectivement les premiers et les derniers textes du corpus. Du point de vue de la ponctuation, les points et points-virgules s'opposent aux virgules et aux points de suspension ; du point de vue des personnes, le *nous* s'oppose au *on* ; du point de vue des temps, le subjonctif présent s'oppose au conditionnel ; enfin, le défini s'oppose à l'indéfini. Ces oppositions semblent pouvoir être regroupées dans une opposition plus générale entre un régime plus assertif, et un régime moins académique,

[1] Classification opérée avec DTM, utilisant la méthode des voisins réciproques.
[2] Au-delà de quatre classes, chaque classe aurait moins de trois individus en moyenne dans l'hypothèse d'une répartition idéale des individus.
[3] On ne garde pour les variables positives et négatives que les variables d'une valeur test supérieure à 2.

notamment par l'indétermination de la personne (*on*), la suspension des phrases (points de suspension), et le conditionnel. Si l'on revient à l'examen du plan factoriel, d'autres variables corroborent cette interprétation : la présence de toutes les marques de structuration (titres, note, division) du côté des textes de la première classe, davantage structurés et organisés.
Les textes de la classe 3 sont faiblement caractérisés. Le plan factoriel, plus riche mais moins sûr, permet de leur associer notamment le point d'exclamation, le *je*, le *tu* et le *vous*. Il oppose fortement le *je* et le *nous*, tandis que les marques de citation sont centrées : cette variation n'est pas due seulement à une variation dans le nombre des citations[1]. L'hétérogénéité des variables utilisées permet peut-être de mieux contrôler l'interprétation d'une variable grâce à une autre.
Les variables utilisées permettent donc de dégager des ensembles de traits en opposition entre les textes corrélés à l'opposition sur l'axe chronologique. Ils ne permettent pas cependant de mener l'analyse à un niveau de précision supérieur à une opposition entre textes académiques (première période), et moins académiques (dernière période). Les textes de la période centrale semblent cependant se détacher à travers des variables, en particulier de personne, qui indiquent un régime dialogique très spécifique.

3. Du texte au paragraphe

La répartition obtenue est-elle similaire si l'on utilise comme individus non plus les textes, mais les paragraphes ? Afin de comparer ces deux paliers, et de prendre en compte plus finement l'hétérogénéité interne des textes, les paragraphes ont été utilisés comme individus, décris par les mêmes variables que précédemment, et soumis à la même analyse factorielle. Les variables hors classe « nombre de paragraphes », « nombre de titres » et « nombre de divisions » ont été retirées, puisque les individus ne contiennent plus ces traits structurels. La variable « nombre de citation » a également été retirée pour des raisons pratiques : cette variable est manquante pour certains textes et la notation de son absence est plus difficile sur des individus nombreux.
Les valeurs propres des premiers facteurs sont, cette fois, beaucoup plus faibles (9%, 8%, 6%, 6%). Les intervalles d'Anderson suggèrent que les premiers facteurs sont peu stables et se recoupent partiellement. Les cinq premiers facteurs au moins semblent nécessaires à une restitution acceptable des variables initiales.

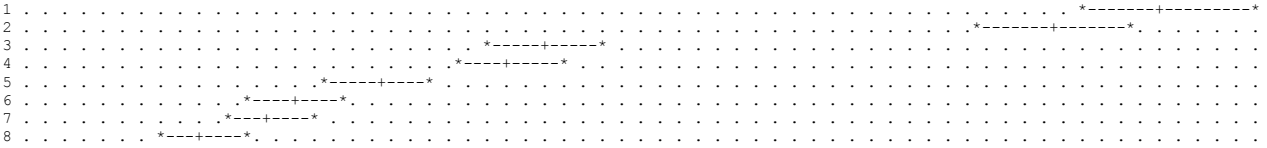


Figure 6 : Représentation des intervalles d'Anderson des cinq premiers axes

Une variable illustrative « titre du texte », placé sur chaque individu, permet de projeter sur le plan des deux premiers facteurs issus de l'analyse des paragraphes les positions respectives des textes correspondants. On observe une structure très proche de la structure précédente : à nouveau, seuls *Spinoza et le problème de l'expression* (SE) et *Logique du sens* (LS) s'opposent à une partition chronologique des textes sur l'origine du premier facteur.

[1] Elle peut naturellement être due à une variation dans leur contenu.

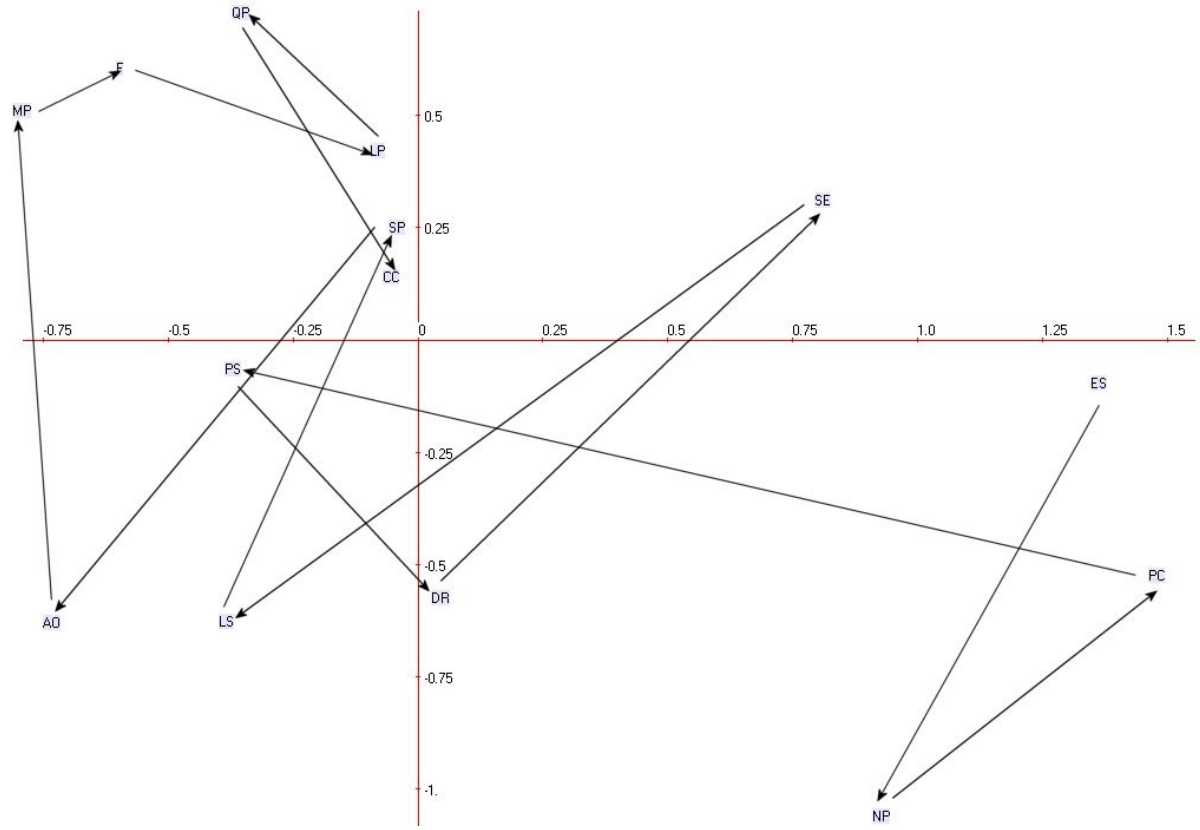


Figure 7 : positionnement de la variable illustrative « texte » sur les deux premiers facteurs de l'analyse en composante principale des paragraphes du corpus

La projection des variables sur les deux premiers facteurs est également proche des distributions observées précédemment.
Les deux premiers facteurs ont cependant une représentativité faible et l'examen de la classification ascendante est nécessaire. On observe une distribution identique des textes à travers les quatre classes, seuls sont écartés cette fois *Proust et les signes* (PS) et *Spinoza philosophie pratique* (SP), les deux commentaires de la classe 4, tandis que, au contraire, *Nietzsche et la philosophie* (NP) réintègre la classe attendue.
L'examen des variables caractéristiques de chaque classe fait apparaître une meilleure distribution des différentes personnes dans les trois classes et, ainsi, une meilleure répartition en ensembles interprétables d'un point de vue dialogique. La quatrième classe acquiert en particulier une caractérisation par la variable *vous*.

Classe		Modalités caractéristiques	Variables caractéristiques positives (personne)	Variables caractéristiques négative (personne)
1	1070	ES PC SE NP	*Nous*	*On*, *ils*, *vous*
2	1171	MP QP F LP	*Ils*, *on*	*Il*, *je*, *nous*
3	3			
4	1386	AO LS DR	*Il*, *vous*	*Nous*, *ils*

Figure 8 : répartition des catégories « noms du textes » dans les quatre classes issues d'une classification ascendante de tous les paragraphes du corpus

C'est cette structure dialogique, opposant les textes entre eux, que nous pouvons maintenant essayer de décrire plus précisément du point de vue de chacun des textes.

4. Comparaison des structures internes des textes

Afin d'affiner la comparaison, on a répété l'expérience au niveau des paragraphes pour trois textes du corpus issus des trois classes observées : *l'Anti-Œdipe*, *Empirisme et subjectivité*, et *Mille*

Plateaux. Les paragraphes de l'*Anti-Œdipe*, répartis en quatre classes[1], font apparaître principalement une opposition structurée autour des personnes :

Classe	Nombre de paragraphes	Variables positives[2]	Variables négatives
1	72	{gram}nom-p, {gram}fin.i, {gram}t.p, {c}, {lm}je	{gram}nom-s, […], {c}!, […],{c}:, {gram}t.subimp, {c}..., {gram}t.subpre
2	51	{gram}t.p, {lm}vous, {gram}fin.d, {lm}tu, {c}? [{c}., {c}:]	{gram}m.partpres, {gram}t.partpas, {c}, {gram}fin.i, {ff}ils
3	94	{gram}t.partpas, {gram}nom-s, {gram}m.partpres, {gram}fin.d, {ff}il, {c}.	{gram}nom-p, {gram}fin.i, {gram}t.p, {gram}ty.p, {c}..., {lm}nous, {gram}t.subimp, {c}!, {lm}vous […] [{ff}ils, {lm}je, {c}?, {lm}tu]
	72	{lm}nous, {c}!, {c}..., {gram}ty.p, […], {gram}t.con, {gram}t.fut, [{tag}hi, {c}?, {c}:]	{ff}il, {gram}t.p, {c}., {gram}nom-p

Figure 9 : classification des paragraphes de l'Anti-Œdipe

La classe 3, la plus nombreuse par le nombre de paragraphes sélectionnés, est remarquable par son exclusion de toute personne autre que la troisième personne du singulier. Ce caractère impersonnel est renforcé par deux temps particulièrement nominaux (les participes présent et passé) et par les noms singuliers. Du point de vue des ponctuations, seul le point est positivement associé, tandis que tous les signes de ponctuation supposables fortement « dialogiques » sont exclus : points d'exclamations, points d'interrogation et points de suspension. Noms communs pluriels et noms propres sont également corrélés négativement.
Chacune des autres classes se spécialise sur une personne : la première du singulier pour la première classe, les secondes personnes du singulier et du pluriel pour la seconde classe, et la première du pluriel pour la dernière. Dans la seconde classe, les secondes personnes sont associées au défini, aux points d'interrogation, et s'opposent aux participes. Dans la quatrième classe, la première personne du pluriel est associée aux points d'exclamation, aux points de suspension, aux noms propres, au futur, au conditionnel, et, dans une moindre mesure, aux points d'interrogation et aux marques de mise en forme. Les variables opposant les paragraphes font donc fortement ressortir une opposition de régimes dialogiques : régime impersonnel (classe 1), régime plus rare (51 occurrences) de l'adresse (classe 2 : présent, seconde personne, points d'exclamation), manifestation du philosophe - incluant éventuellement ses destinataires - (classe 4 : 1ère personne du singulier, points d'exclamation et de suspension, temps modaux), et un régime dialogique moins caractérisable (première personne du singulier, noms propres, présent, virgule).
Empirisme et subjectivité, opposé à l'*Anti-Œdipe* par le premier facteur de l'analyse précédente, montre une structure différente[3].

[1] Les valeurs propres des facteurs sont là encore relativement faibles : 11%, 9%, 7% et 6% pour les quatre premiers. Les intervalles d'Anderson sont par contre faiblement recouverts.
[2] Au-dessus d'une valeur test de 2. Des variables au-dessous de ce seuil sont notées entre crochets droits. Certaines variables sont exclues (noté […]).
[3] Les quatre premières valeurs propres sont plus faibles que celles de l'analyse de l'*Anti-Œdipe* : 9%, 8%, 7% et 5%.

Classe	Nombre de paragraphes	Variables positives[1]	Variables négatives
1	76	{gram}t.p, {gram}nom-s, {gram}t.fut , {gram}t.con, […] {lm}on […]	{gram}m.partpres, {gram}t.partpas, {gram}nom-p, {ff}ils […] {lm}nous
2	2	{lm}vous, {c}?, {tag}note, {lm}je	{c}.
3	77	{gram}m.partpres, {gram}t.partpas, {gram}fin.d, {gram}nom-s	{gram}t.p, {gram}fin.i
	76	{gram}nom-p, {ff}ils, {c};, {c}: {gram}ty.p, […]	{gram}t.con, {gram}t.fut, {tag}hi […]

Figure 10 : classification des paragraphes de Empirisme et subjectivité

Ici, il est délicat d'interpréter les spécificités des classes comme des ensembles cohérents d'un point de vue dialogique. Les seules personnes présentes, si l'on met à part la troisième personne, sont dans une même classe ne comptant que deux individus (classe 2). Le présent, qui s'opposait précédemment au futur et au conditionnel, leur est maintenant corrélé. Seule la première classe peut être interprétable, le *on* remplaçant le *nous* dans le rôle de la manifestation du philosophe.

Classe	Nombre de paragraphes	Variables positives[2]	Variables négatives
1	277	{gram}fin.i, {gram}nom-p, {gram}t.p, {lm}on, {lm}nous	{gram}fin.d, {gram}nom-s, {gram}ty.p, {gram}t.i, {ff}il
2	162	{gram}fin.d, {gram}nom-s, {gram}ty.p, {gram}m.partpres, {gram}t.partpas, {c}:	{gram}fin.i, {gram}nom-p, {lm}on
3	14	{gram}t.i, {gram}t.subimp, {gram}t.pas, {ff}il	{gram}t.p, {tag}note, {lm}nous, {gram}t.fut
	1		

Figure 11 : classification des paragraphes de Mille Plateau

Les très faibles effectifs de la classe 3 rendent son interprétation délicate. L'opposition principale s'établit entre les classes 1 et 2 : on relève en particulier, comme caractéristique de la première, le présent, le *on* et le *nous*, auxquels s'opposent dans la seconde les noms propres, et les participes. On retrouve dans l'association entre les participes, le défini, et les noms singuliers, une configuration de traits associée à l'absence de marques dialogiques fortes. La principale structure observable à partir de cette classification est l'opposition sur le critère de la présence de marques dialogiques.

Conclusion

Les trois expériences menées ont fait varier les paliers, du texte au paragraphe. Sur le même jeu de variables, on observe à la fois une classification du corpus restituant l'axe diachronique, et caractérisé par une opposition sur le registre académique, et une opposition des textes par la distribution de leurs paragraphes, où peut se lire la marque d'une organisation dialogique.

BIBLIOGRAPHIE

DESCOMBES, V. 1979. *Le même et l'autre : Quarante-cinq ans de philosophie française* (1933-1978), Minuit, Paris.

LEBART, L., MORINEAU, A., PIRON, M. 2000. *Statistique exploratoire multidimensionnelle* (troisième édition), Paris, Dunod.

LEBART, L. 2006. *Data and Text Mining* (DTM) <http://www.lebart.org>

[1] Au-dessus d'une valeur test de 2. Des variables au-dessous de ce seuil sont notées entre crochets droits. Certaines variables sont exclues (noté […]).

[2] Au-dessus d'une valeur test de 2. Des variables au-dessous de ce seuil sont notées entre crochets droits. Certaines variables sont exclues (noté […]).

LOISEAU, S. 2005. Thématique et sémantique contextuelle d'un concept philosophique, in G. Williams (éd.), *La linguistique de corpus*, Rennes^, Presses Universitaires de Rennes, pp. 129140.
LOISEAU, S., POUDAT, C., ABLALI, D. 2006. Exploration contrastive de trois corpus de sciences humaines, *JADT 2006*, Besançon, 19-21 avril 2006.
LOISEAU, S. 2006. CorpusReader, un logiciel d'interrogation de corpus (<http://panini.u-paris10.fr/CR>)
POUDAT, C. 2006. *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*, Thèse de doctorat, Université d'Orléans.
RASTIER, F. 2001. L'Être naquit dans le langage : Un aspect de la mimésis philosophique, *Methodos*, vol. I, n. 1, Presse du septentrion, Lille, pp. 103-132.
RASTIER, F. 2005. Enjeux épistémologiques de la linguistique de corpus, in G. Williams (éd.), *La linguistique de corpus : Actes des deuxièmes Journées de la linguistique de corpus*, Rennes, Presses Universitaires de Rennes, pp. 31-47.
VALETTE, F. 2003. Conceptualisation and Evolution of Concepts. The example of French Linguist Gustave Guillaume, in K. Fløttum & F. Rastier (éds.), *Academic discourse -- multidisciplinary approaches*, Oslo, Novus, pp. 55-74.

RÉFÉRENCES DES TEXTES DU CORPUS

[ES] Deleuze G. (1953 [1998]), *Empirisme et subjectivité*, Paris, PUF.
[NP] Deleuze G. (1962), *Nietzsche et la philosophie*, Paris, PUF.
[PC] Deleuze G. (1963 [1994]), *La philosophie critique de Kant*, Paris, PUF.
[PS] Deleuze G. (1964 [1996]), *Proust et les signes*, Paris, PUF, coll : Quadrige.
[SE] Deleuze G. (1968), *Spinoza et le problème de l'expression*, Paris, Minuit.
[DR] Deleuze G. (1968 [2000]), *Différence et répétition*, Paris, PUF.
[LS] Deleuze G. (1969), *Logique du sens*, Paris, Les éditions de Minuit, coll : Critique.
[SP] Deleuze G. (1970 [1981]), *Spinoza Philosophie pratique*, Paris, Minuit.
[AO] Deleuze G., Guattari F. (1972), *L'Anti-Œdipe*, Paris, Minuit.
[MP] Deleuze G., Guattari F. (1980), *Mille Plateaux*, Paris, Minuit.
[F] Deleuze G. (1986), *Foucault*, Paris, Minuit.
[LP] Deleuze G. (1988), *Le pli – Leibniz et le baroque*, Paris, Minuit.
[QP] Deleuze G., Guattari F. (1991), *Qu'est-ce que la philosophie ?*, Paris, Minuit.
[CC] Deleuze G. (1993), *Critique et clinique*, Paris, Minuit.

ANALYSER LA PRESSE ANCIENNE AVEC LE PROGICIEL *PHPRESS* : LE TRAITEMENT NUMÉRIQUE DES FAITS DIVERS DE *L'ECLAIREUR DE NICE*, 1928-1929

Matthieu PEREZ
Doctorant, Centre de la Méditerranée Moderne et Contemporaine (CMMC), Université de Nice - Sophia Antipolis

SOMMAIRE

Résumé : PhPress *est un progiciel de Gestion Électronique de Documentation conçu pour assister l'historien dans le traitement de la presse ancienne. Développé dans un esprit OpenSource, il s'agit d'un outil puissant permettant la numérisation de vastes fonds de presse quotidienne, la consultation aisée de ces fonds, ainsi que la construction de bases analytiques destinées à l'étude du contenu et de la morphologie des articles. Nous avons construit cet outil en fonction des besoins que nous avions définis au cours de l'élaboration de notre projet de thèse : « La chronique des faits divers dans la presse niçoise des années de crise : 1929-1939 » (thèse en cours de réalisation, sous la direction de M. Ralph SCHOR).*
Nous chercherons dans un premier temps à présenter les raisons de la construction de ce logiciel, avant d'en aborder le fonctionnement concret dans le cadre d'une étape de notre travail de recherche, l'analyse de la chronique des suicides dans le journal régional L'Eclaireur de Nice *de l'année 1928, constituée de deux temps : d'une part la constitution du corpus numérique, d'autre part l'exploitation de ce corpus.*

1. Les raisons d'être de *PhPress*

La construction d'un logiciel dans le cadre d'un travail de recherche en sciences humaines n'est pas très fréquente et il peut sembler nécessaire d'expliquer ce qui a motivé ce travail très lourd et assez inhabituel.

1.1. Brève historiographie du fait divers

Le fait divers est une rubrique de presse constituée de récits d'événements anormaux prenant place dans le cadre de la vie quotidienne : crimes, délits, accidents, anomalies naturelles, faits curieux de toute sorte. Ce sont des récits généralement considérés comme insignifiants, vulgaires, médiocres. Ils occupent pourtant dans la presse une place considérable, particulièrement dans la presse antérieure aux années 1960, et entretiennent avec la littérature une relation très forte : beaucoup d'écrivains ont débuté leur carrière comme fait-diversiers, beaucoup de livres importants ont pour origine un fait divers.

C'est dans le champ littéraire que les bases de l'analyse du fait divers vont être posées : Roland Barthes en définit la structure narrative en 1964. Puis Georges Auclair l'observe sous un angle anthropologique en 1970.
On a longtemps considéré ce genre de récit comme non historique : le fait divers apparaissait comme un objet littéraire ou anthropologique, mais ne trouvait place que dans une « petite histoire » anecdotique opposée à la « grande histoire ». Observé sous l'angle de l'anecdote, le fait divers apparaissait avant tout comme un divertissement, et beaucoup de titres consacrés à ce sujet sont d'abord ludiques ; l'historien utilisait ce matériau avant tout comme simple illustration.
Avec Dominique Kalifa et ses recherches sur le récit de crime à la belle-époque, qui présente le fait divers et la littérature populaire comme un ensemble de récits normalisateurs, l'appréhension du fait divers par l'historien prend un nouvel aspect.
Deux thèses entièrement consacrées au fait divers vont immédiatement suivre, avec Marine M'Sili et Anne-Claude Ambroise-Rendu, qui abordent le fait divers à travers des méthodes nouvelles : le comptage des articles, des références géographiques qu'ils contiennent, la ventilation des articles en fonction de leur contenu, prennent une importance centrale dans le traitement des sources, dont l'interprétation se fait largement au moyen de méthodes statistiques. Ces nouvelles méthodes correspondent à une prise en compte du récit de fait divers comme un objet historique à part entière, où le contenu, la répétition, la sur ou sous-représentation de certains thèmes, vont être porteurs de sens.
Le travail que nous réalisons actuellement vise à étudier, dans le contexte très particulier de la Riviera durant les dix années précédant la seconde guerre mondiale, le comportement de la chronique des faits divers de deux quotidiens locaux, *L'Eclaireur de Nice* et *Le Petit Niçois*, représentant deux tendances politiques opposées.

1.2. Les obstacles techniques limitant les champs de recherche de ce domaine

La rubrique des faits divers occupe généralement une très grande place dans la presse ancienne, et on en trouve habituellement 30 à 60 articles par numéro dans les quotidiens généralistes. Cela représente très vite, même sur une période relativement courte, une masse documentaire considérable, dont la simple manipulation devient problématique.
Tant que le traitement des documents reste purement manuel, cette surabondance de matière, indissociable de cette nouvelle approche méthodologique plaçant le fait divers et son contenu au centre de la problématique, oblige les chercheurs à procéder à des sondages très dilués sur leur corpus, et l'on perd ainsi énormément d'information : par exemple, le rythme saisonnier de la chronique va être masqué par un sondage ne portant que sur une courte partie de l'année. Par ailleurs, un simple comptage des articles, s'il permet l'élaboration de tableaux statistiques utiles, ne permet pas de les hiérarchiser et de faire la différence entre une information importante et une simple dépêche.
Les méthodes d'analyse morphologique de la presse quotidienne définies par Jacques Kayser permettraient de lever ce genre d'ambiguïtés, mais elles sont difficilement utilisables sur un corpus important : mesurer la surface de quelques articles dans quelques numéros est aisé, le faire pour plusieurs centaines ou plusieurs milliers d'articles est impossible, du moins dans des délais acceptables.
Enfin, il est très difficile de revenir en arrière et de retrouver un article particulier ou mal saisi lorsque l'on travaille de façon traditionnelle sur un corpus de presse quotidienne : copier tous les articles étudiés serait trop coûteux, trop long, et poserait le problème de la gestion d'une telle masse de documents ; la manipulation des microfilms ou des liasses d'archive est suffisamment compliquée pour que l'on ne revienne à un document déjà consulté qu'en cas de très grande nécessité, et que l'on ne travaille habituellement qu'à partir de notes, prises parfois un peu vite. Ces difficultés, observées concrètement lors de la réalisation de notre mémoire de maîtrise, nous semblaient très handicapantes, et nous avons recherché une solution technique pour les dépasser.

1.3. Les solutions théoriques pour y remédier

Nous avons cherché le moyen d'optimiser autant que possible le travail du chercheur, de façon à augmenter le volume de documents et de données manipulés dans un temps aussi court que possible, afin de rendre possible un travail sur le fait divers mettant en œuvre un corpus vaste, accroissant la précision de la recherche et la fiabilité des statistiques obtenues. Il fallait donc un

logiciel adapté. Les contacts que nous avons eus avec les services informatiques universitaires et avec différentes entreprises professionnelles de la Gestion Électronique de Documentation se sont vite avérés décevants : si de nombreux systèmes de gestion de documentation existaient, aucun ne permettait de faire efficacement le travail que nous projetions.
Nous avons donc déterminé un cahier des charges pour une application informatique, qui devrait contenir une base de données contenant l'ensemble de l'analyse de contenu appliquée aux faits divers, optimisant ainsi le travail statistique classique réalisé sur cette rubrique, et qui permettrait également de mener une analyse morphologique des articles, ainsi que de les conserver sous forme électronique, pour avoir la possibilité de les retrouver et de les consulter aisément.
Dans tous les cas, il nous semblait primordial de réduire autant que possible les gestes à accomplir pour utiliser le logiciel, et d'optimiser au maximum la productivité du système.
D'autre part, il nous a semblé utile de chercher à réaliser un logiciel qui serait un logiciel libre, pour que son développement puisse éventuellement se poursuivre dans le cadre de la communauté Open-Source, pour que sa diffusion puisse se faire sans contrainte, et surtout pour en rendre visibles les mécanismes : en théorie, chacun peut ainsi vérifier que le logiciel fonctionne réellement et ne donne pas de résultat erroné ; d'un point de vue pratique, le choix du développement Open-Source permet également de faire appel à de nombreux logiciels libres existants, et de les inclure dans le fonctionnement du système – ce qui ne serait pas possible dans une logique de développement propriétaire.
Nous voulions également que le logiciel soit aussi modulaire que possible, et puisse servir à d'autres recherches que la nôtre.
Enfin, nous voulions garantir une certaine interopérabilité, et faire en sorte que le système soit utilisable depuis des ordinateurs fonctionnant avec des applications Microsoft, mais également sous Linux et Mac OS.
À partir de ce cahier des charges, des choix techniques précis se sont imposés de façon assez automatique.

1.4. La mise en application de ces solutions et la construction de *PhPress*

Parmi les diverses solutions techniques disponibles, les différents langages de programmation, le choix d'un système basé sur PHP et MySQL s'est rapidement imposé : il s'agit de techniques Open-Source, très populaires et donc bien documentées, d'un abord relativement facile ; PHP est un préprocesseur Hypertexte, qui génère automatiquement des pages Web, et qui dispose de nombreuses fonctions de traitement graphique et logique ; il permet de construire facilement des interfaces vers un système de gestion de bases de données relationnelles tel que MySQL, un logiciel libre, assez facile d'accès et très répandu.
Nous avons donc appris les bases de la programmation PHP, et commencé la construction du logiciel – construction encore inachevée, puisque bien que déjà utilisable, *PhPress* ne dispose pas encore de toutes les fonctionnalités que nous avons prévues.
Techniquement, *PhPress* fonctionne comme un site Web dynamique ; il met en oeuvre un ordinateur employé comme serveur, qui doit obligatoirement fonctionner sous Linux, qui contient la base de données et les images représentant les pages de journal numérisées, et un ordinateur client, utilisé pour se connecter au serveur, qui doit, pour présenter de bons résultats, être un Macintosh disposant de Mac OSX et du navigateur Firefox. L'interopérabilité que nous recherchions n'est pas encore atteinte, mais reste un objectif à long terme – notre priorité étant de réaliser notre propre travail de recherche.

2. La constitution du corpus documentaire

2.1. L'acquisition numérique des documents depuis les sources originales ou microfilmées

Au début de notre travail, la BNF ne prévoyait pas encore de numériser ses collections de journaux anciens, et les archives départementales des Alpes-Maritimes ne pouvaient pas avancer de date ni donner de détail technique concernant leur campagne de numérisation de la presse quotidienne locale. Nous avons donc choisi, en accord avec la direction des Archives Départementales, de numériser nous-mêmes, à partir d'un lecteur-numériseur de microfilms, les collections dont nous avions besoin, en commençant par l'année 1928.
Nous avons donc équipé *PhPress* d'un dispositif permettant d'indexer manuellement des images au format TIFF (le seul format d'image possible avec le matériel dont nous disposions) au fur et à mesure de la numérisation d'un rouleau de microfilm, et nous avons ainsi numérisé la collection

complète du principal quotidien régional de la période, *L'Eclaireur de Nice*, entre 1928 et 1932, soit 12 214 pages ; nous avons construit des outils permettant de gérer la sauvegarde sur CDROM de ces documents et de les convertir dans un format manipulable par PHP, le format PNG.
Depuis, les Archives Départementales des Alpes-Maritimes ont achevé une tranche importante de leur propre campagne de numérisation de leurs collections, qui incluent *L'Eclaireur de Nice* et *Le Petit Niçois* ; ces journaux ont été numérisés depuis les pages originales, ce qui permet une qualité bien supérieure à celle obtenue par numérisation des microfilms. Ces documents ont été mis en ligne et sont accessibles depuis le site des Archives Départementales, mais avec un temps de réponse très long du serveur et une interface qui ne se prête pas à une exploitation automatique, ce qui rend impossible l'exploitation systématique de ces pages ; la direction des archives a accepté de nous transmettre une copie de ces documents, mais la constitution de cette copie semble poser des problèmes et nous restons pour l'instant dans l'expectative. Nous ignorons encore les spécificités techniques de ces images, et nous ne pourrons donc construire un module permettant de les intégrer à *PhPress* que lorsque cette situation sera débloquée.

2.2. Repérage et sélection des articles de faits divers

Les images numériques représentant les pages de journal doivent, avant d'être utilisées, subir un étalonnage permettant de connaître leur surface réelle, de façon à compenser les différences d'échelle dues à la focale optique du numériseur de microfilm. Nous avons ainsi réalisé une mesure de la surface imprimée réelle des pages de *L'Eclaireur de Nice* pour la période qui nous concerne – période où le format du journal reste assez stable – qui permet de convertir approximativement en millimètres les surfaces en pixels des images informatiques.
Pour chaque page, l'utilisateur doit donc désigner d'un clic les quatre points extrêmes de la surface imprimée de la page affichée ; à partir des coordonnées de ces points, l'ordinateur calculera la surface relative de chaque article qui sera sélectionné dans cette page, sélection qui se fait également d'un clic sur les quatre points extrêmes de la surface choisie. À partir des coordonnées de ces points, l'ordinateur calculera la surface de chaque article et estimera sa position dans la page de journal. La sélection des articles se fait également par sélection à la souris des quatre coins du bloc de texte ou d'illustration. Un masque de couleur transparente est appliqué sur le document de façon à rendre visible le travail réalisé, et de distinguer les différents types d'articles présents sur la page.

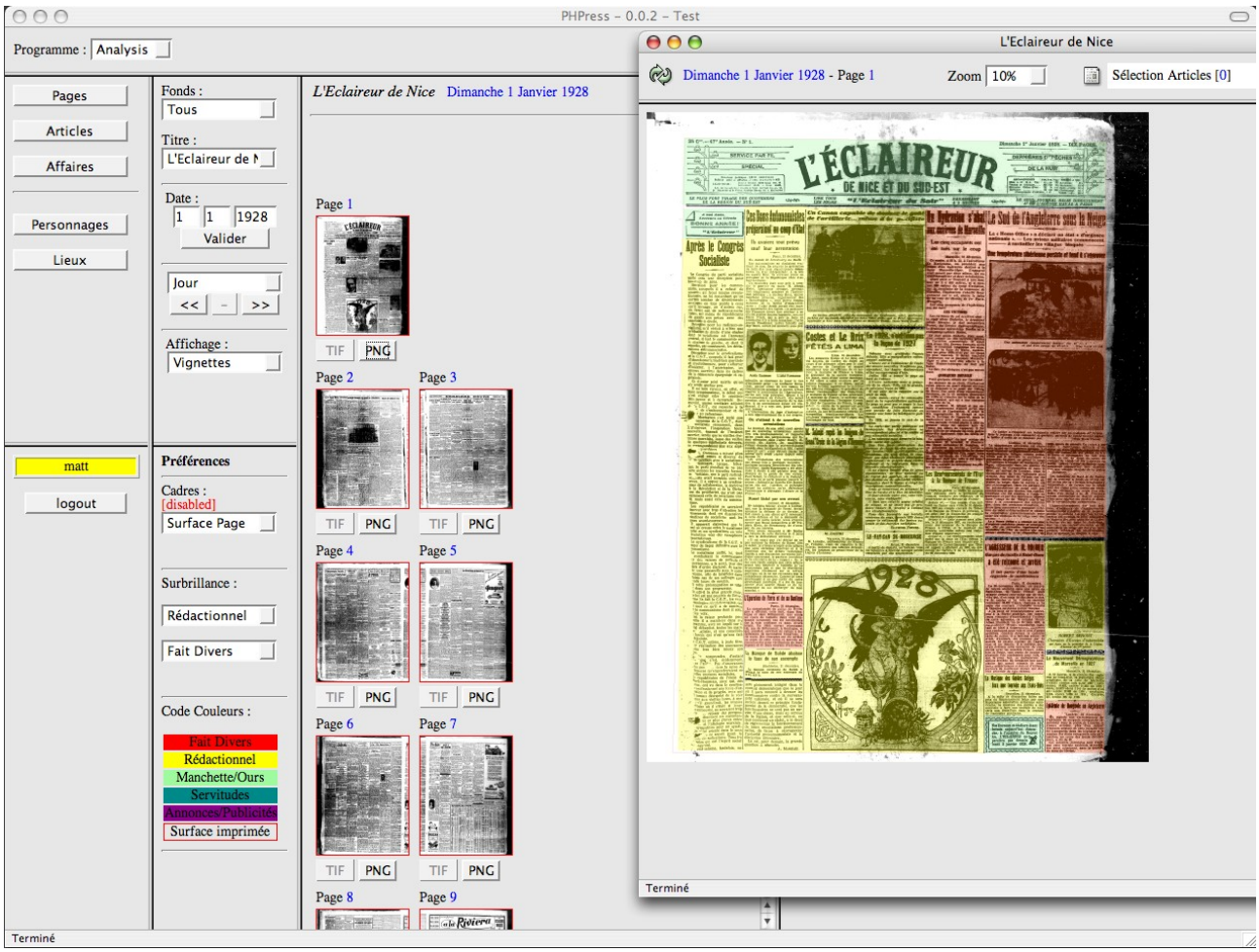


Illustration 1 : exemple de page de journal traitée avec PhPress

2.3. L'étiquetage simple des articles : titre, thématique principale, géographie, morphologie

Une fois un article sélectionné, un masque apparaît automatiquement et demande à l'utilisateur de saisir les informations essentielles permettant d'indexer ce document : son titre, sa nature (article rédactionnel, rubrique de servitudes, publicité, élément de titre), sa typologie (article de fait divers, de politique, etc.) ainsi que la sous-catégorie dans laquelle on peut le placer : on pourra ainsi ventiler les faits divers entre récits de suicide, récits de crime, etc., et analyser séparément ces différents genres de faits divers.

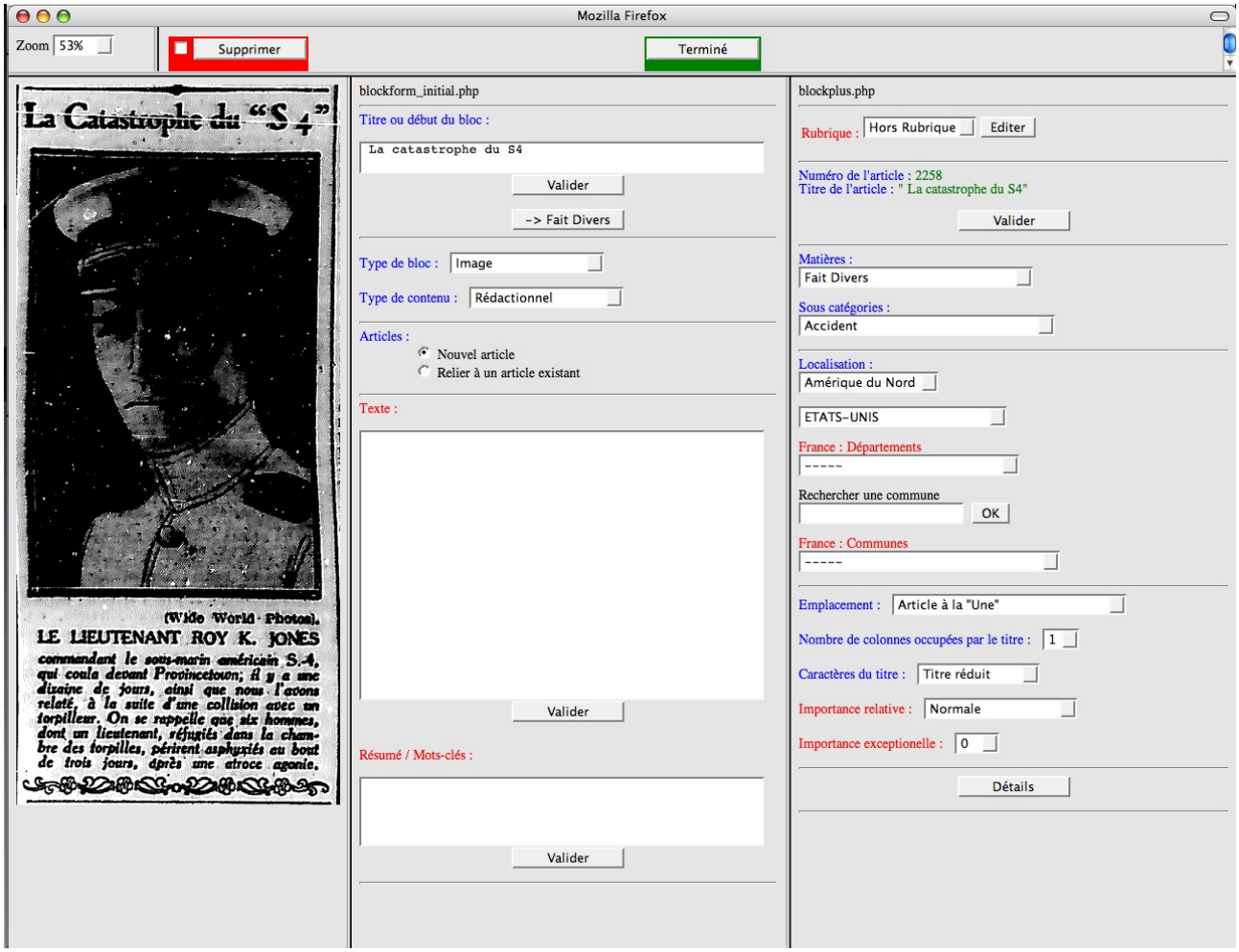


Illustration 2 : un exemple de masque d'indexation des articles

On peut relier chaque article à une indication géographique plus ou moins précise : continent, pays, et, pour la France, département et commune. Les données géographiques sont calquées sur les bases de données de l'INSEE.
Le système prévoit également un formulaire permettant de calculer un indice de mise en valeur inspiré par la méthode de Jacques Kayser et partiellement automatisée, qui permet de hiérarchiser les articles en fonction de leur aspect graphique. Cet indice est complémentaire d'une autre information essentielle calculée de façon automatique par le logiciel, la surface de l'article – exprimée en pourcentage de la surface imprimée de la page :

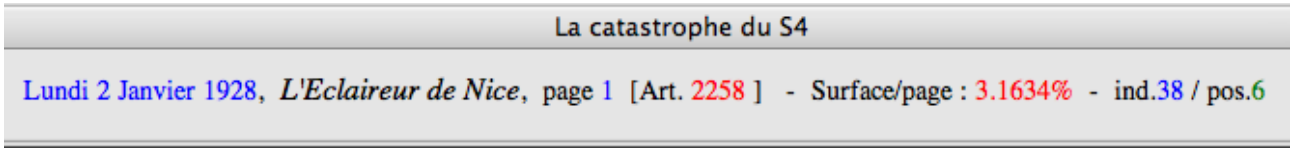


Illustration 3 : affichage de la surface relative de l'article, de l'indice de mise en valeur et de la position de l'article

L'indication de position de l'article est un chiffre de 1 à 9 correspondant à neuf secteurs de la page de journal, calculé automatiquement à partir des coordonnées du coin supérieur gauche de l'article.

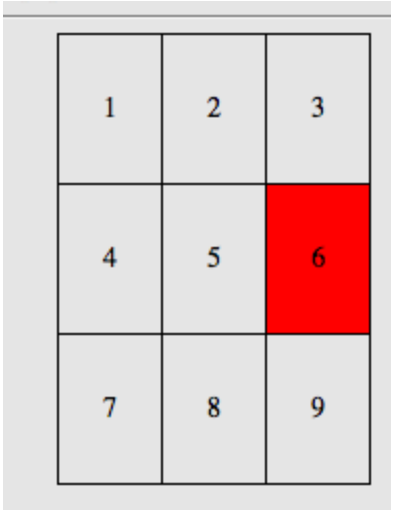


Illustration 4 : repérage de la position de l'article dans la page

Une fois indexé, l'article est accessible depuis l'affichage analytique correspondant au numéro du journal :

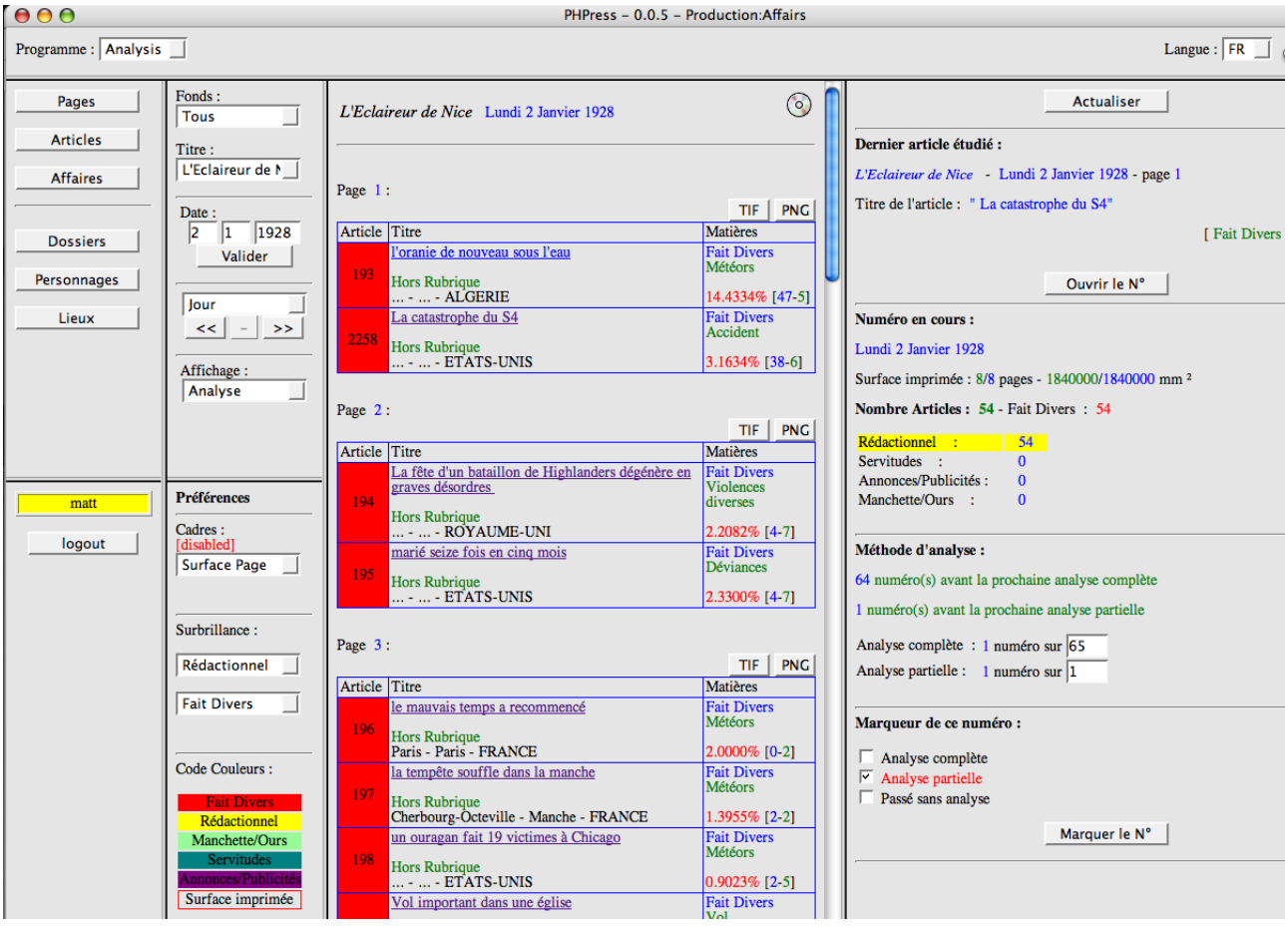


Illustration 5 : affichage analytique du numéro du 2 janvier 1928 de L'Eclaireur de Nice

Les articles, après ce premier marquage, peuvent déjà être triés et comptés en fonction de leur position, de leur taille, de leur typologie, des lieux géographiques mentionnés. Un étiquetage plus fin est également possible.

2.4. L'étiquetage avancé des articles

Les articles s'affichent dans une fenêtre mettant en regard trois cadres, d'abord l'article lui-même, le cadre central contenant un résumé des informations contenues dans l'article, le cadre de droite permettant l'édition de ces informations :

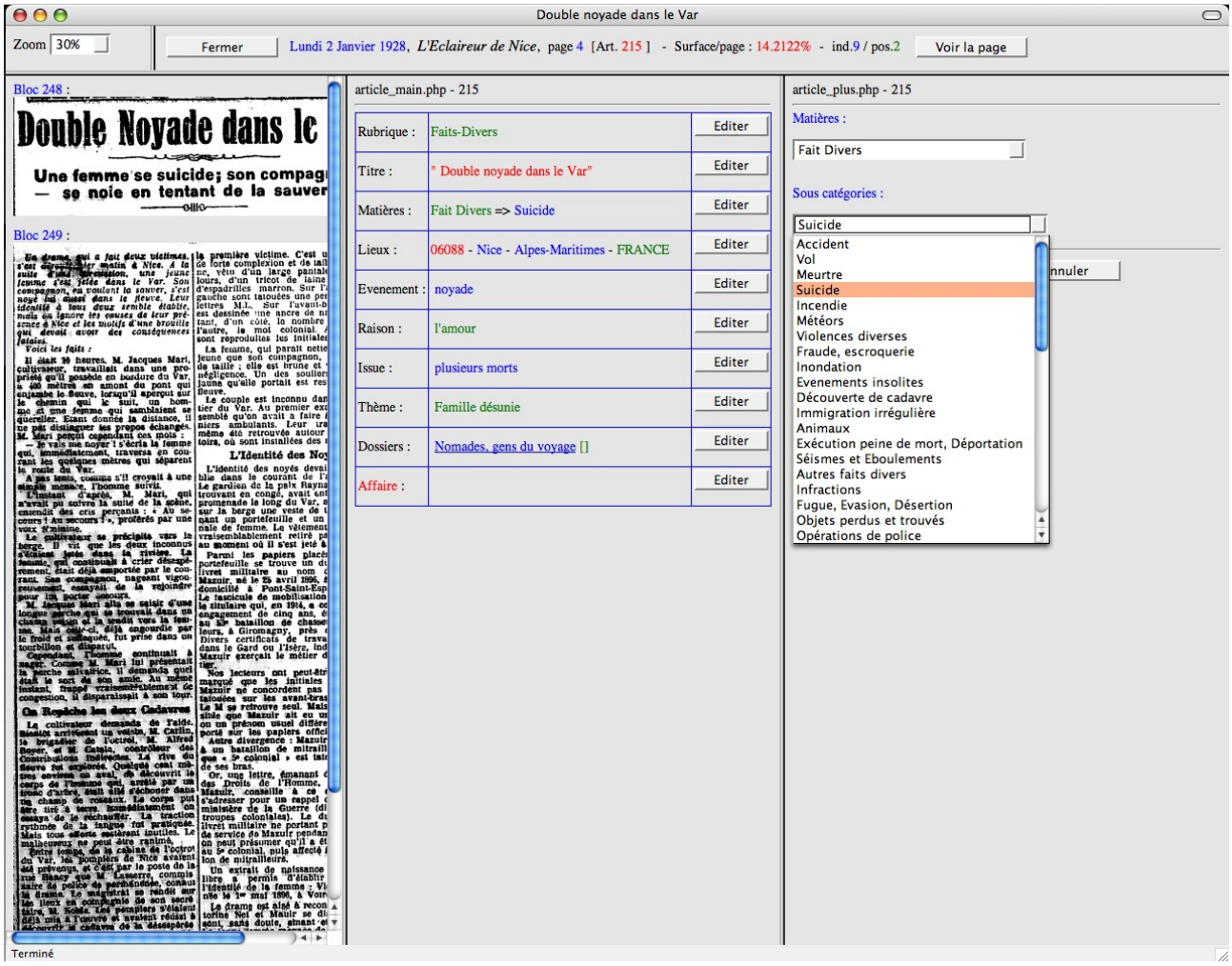


Illustration 6 : exemple d'affichage d'un article et édition des marqueurs qui le caractérisent

Nous pouvons ainsi ajouter à la description sommaire de l'article réalisée lors de sa sélection des marqueurs plus fins, tels la nature précise de l'événement, les motifs des acteurs, l'issue, le thème principal de l'article. Une fonction permettant d'insérer l'article dans une cartographie précise, tenant compte du nom des rues et permettant de les repérer sur un plan, est en cours d'élaboration.

Il est possible de construire un certain nombre de dossiers, et de relier ainsi les articles présentant des caractéristiques communes, déterminées par le chercheur.

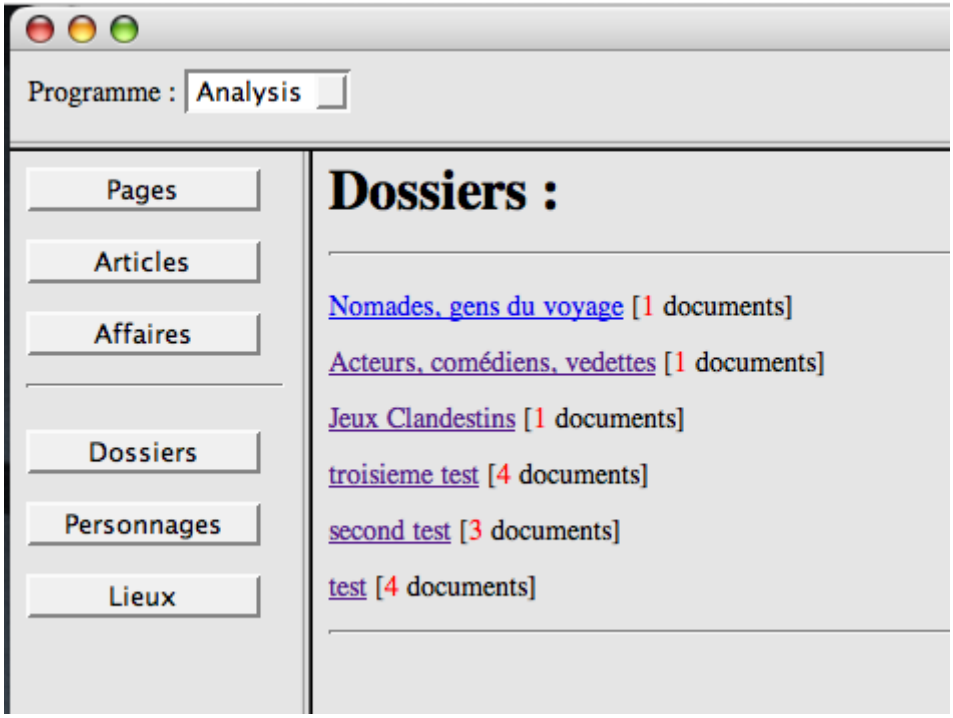


Illustration 7 : liste de dossiers

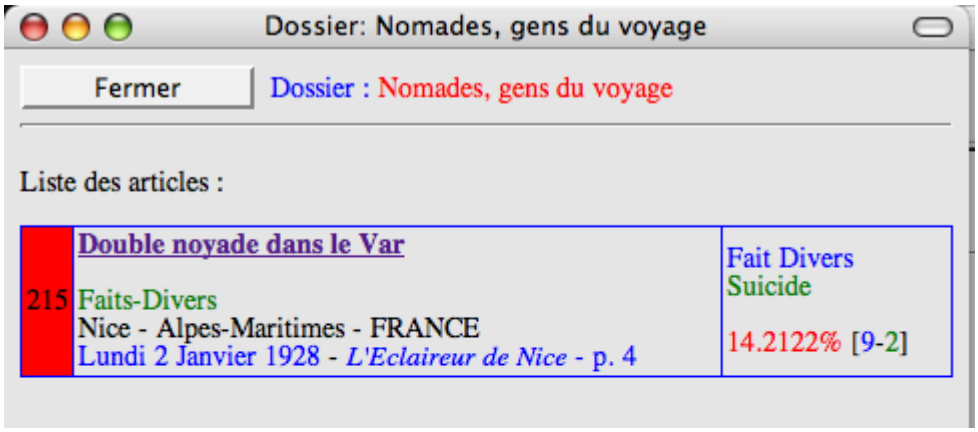


Illustration 8 : un exemple de dossier

La base de données comportera prochainement une table permettant de suivre les articles en tant qu'affaires, c'est-à-dire de repérer les articles se suivant et concernant un même événement, ainsi que les articles équivalents publiés par plusieurs journaux. Une autre table permettra de constituer un fichier des personnes apparaissant dans les articles, et de noter les actions et les relations entre ces personnes. Cependant, nous rencontrons des difficultés techniques dans la réalisation de ces deux éléments, dont le développement a pris beaucoup de retard. D'une façon générale, l'étiquetage avancé des articles est encore en période de test, seul l'étiquetage simple ayant été utilisé sur un volume conséquent de documents.

3. Exploitation de la base et interprétation des résultats

Nous avons étiqueté 2 468 articles de faits divers parus dans *L'Eclaireur de Nice* durant l'année 1928, dont 217 articles représentant des suicides. À partir de cette documentation réduite, nous avons pu faire un test de la productivité de notre système et des contraintes pesant sur son utilisation, qui a entraîné un certain nombre de modifications dans le logiciel. Les routines permettant l'exploitation de la base de données restent pour l'instant très sommaires et manquent de maturité, mais permettent tout de même d'obtenir certains résultats.

3.1. Axe morphologique

La prise en compte de la morphologie des articles est un point très important de notre système. Pour le moment, l'exploitation reste assez limitée et se fait au travers d'un formulaire permettant de trier les articles par thème, par date, par surface et par indice de mise en valeur :

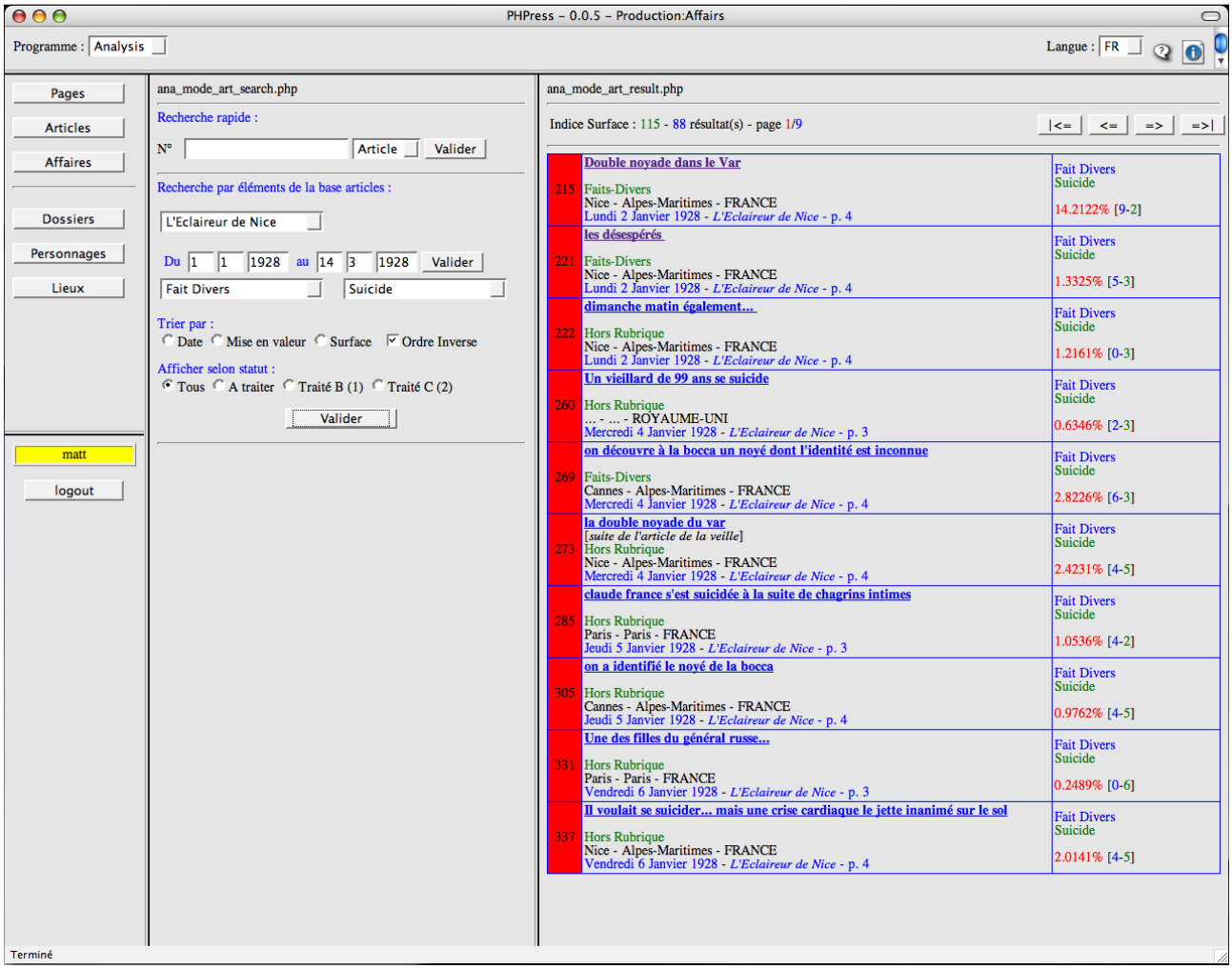


Illustration 9 : recherche d'articles de suicide en fonction de leur date de publication

Le système, en fonction de ces critères de recherche, fournit une estimation de la surface réelle cumulée des articles répondant aux critères ; nous avons ainsi pu élaborer des tableaux présentant l'évolution sur un an des récits de suicide dans *L'Éclaireur de Nice.*

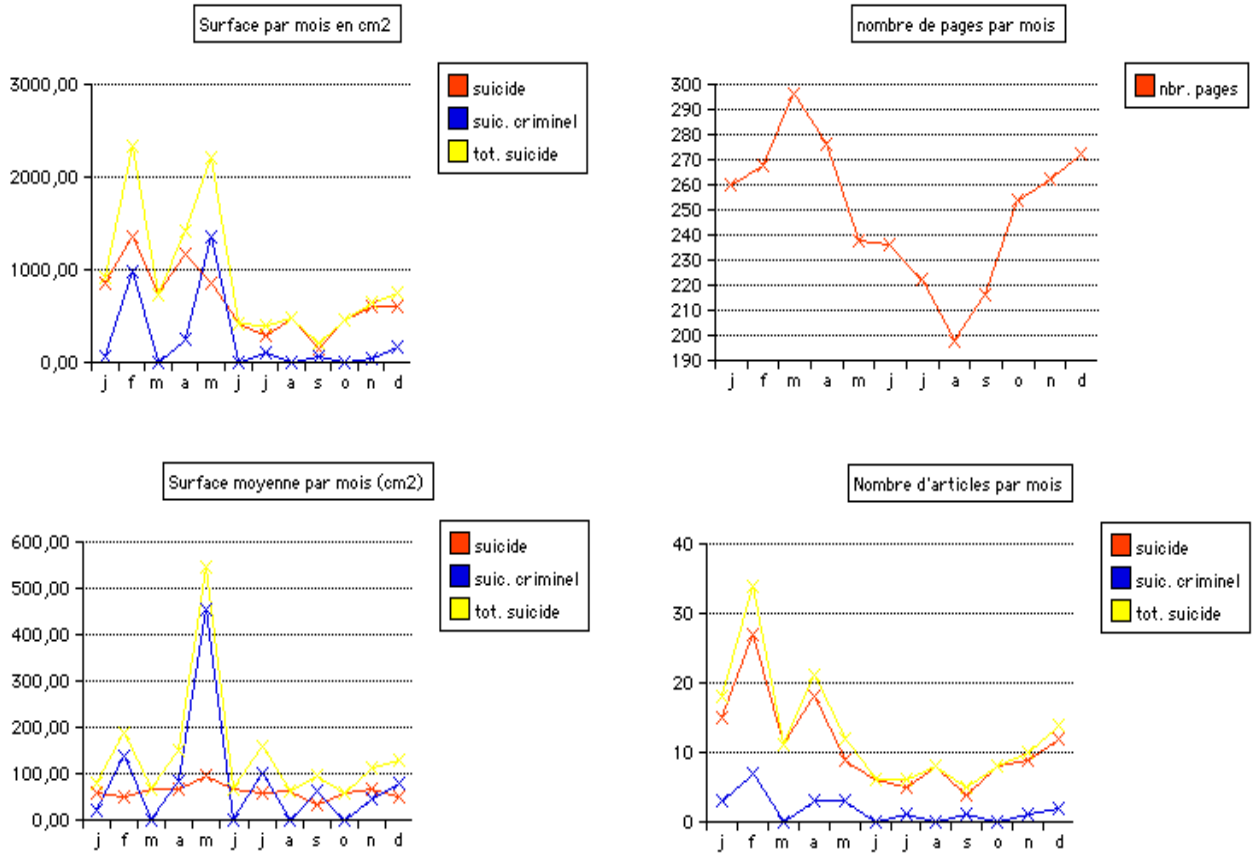


Illustration 10 : évolution des récits de suicide au cours de l'année 1928 dans L'Eclaireur de Nice

L'ensemble du système reste très sommaire et appelle de nouveaux développements, qui sont en cours de réalisation.

3.2. Axe géographique

Nous n'avons pas encore achevé les scripts permettant d'interroger la base de données à partir de critères strictement géographiques, et nous n'avons donc pas encore pu utiliser pleinement cette information. Les données géographiques concernant les articles sont libellées dans les mêmes normes que celles de l'INSEE. Notre objectif à court terme est de permettre un tri simple des articles à partir de critères géographiques ; dans un deuxième temps, la représentation des données sous forme de carte est envisagée, mais nous manquons de temps pour réaliser le développement de ces applications.
Une géographie fine, permettant d'enregistrer le nom des rues où se déroulent les événements, est déjà en place mais n'a pas encore été utilisée à grande échelle.
L'inscription dans l'espace géographique des récits de fait divers est un point essentiel de notre travail de recherche, nous sommes donc très mobilisé par cette question.

3.3. Taxinomies

L'étiquetage des articles se fait avant tout par l'insertion des articles dans une classification : classement par catégories d'articles (politique, faits divers, etc.) et de là en sous-catégories (crime, suicide, sous-catégories de l'ensemble fait divers, ...). Mais concrètement, les articles de presse sont parfois difficiles à classer dans une catégorie trop précise ; la multiplication des cas particuliers ou limites nous a dans un premier temps poussé à multiplier les sous-catégories rares, avant de développer des marqueurs supplémentaires permettant de limiter les rubriques de classement principal, sans perdre les particularités des articles. L'utilisation à grande échelle de *PhPress* nous a poussé à modifier complètement la liste des sous-catégories de faits divers que nous avions établie à l'avance, et qui ne correspondait pas suffisamment à la réalité des articles. Les catégories permettant de marquer les autres articles, que nous avions construites en nous

fiant un peu servilement à une liste de rubriques présentée par Jacques Kayser, se sont révélées pratiquement inutilisables, et doivent être complètement révisées.

3.4. Exploitation avancée des données

PhPress est un système dont le développement est en cours, et qui souffre donc de nombreuses lacunes ; l'exploitation des données reste particulièrement en retard sur d'autres fonctions complexes, dont l'élaboration a pris beaucoup de temps. Nous ne pouvons pour l'instant présenter que des fonctionnalités qui sont en cours d'élaboration ou de conception.

Le principal axe de développement du logiciel est le traitement des séries d'articles, des affaires, qui permettra de déterminer comment des journaux concurrents traitent des mêmes sujets ; il s'agit là d'un point particulièrement important, mais également particulièrement complexe, que nous sommes actuellement en train de mettre au point.

D'autre part, le traitement des personnages représentés fournira des clés importantes, mais demande une réflexion technique importante qui n'est pas achevée.

En conclusion, *PhPress* est un logiciel qui reste inachevé, et qui évolue au cours de l'utilisation que nous en faisons ; il est clair qu'il n'est pas encore pleinement fonctionnel, et qu'il n'est pas utilisable sans un important apprentissage. Construit par un historien, dans un but précis et entièrement déterminé par nos objectifs de recherche, il remplit son rôle essentiel correctement et représente un outil augmentant considérablement la puissance de travail de l'historien. Tant que son utilisateur reste son concepteur, l'outil se plie naturellement aux impératifs du chercheur ; mais, publié, il sera certainement difficilement maîtrisable pour les utilisateurs, même de bon niveau ; il reste donc un très grand travail de développement à réaliser, qui devra viser à rendre l'ensemble du système de *PhPress* compréhensible et aisément modifiable par les utilisateurs.

BIBLIOGRAPHIE

AMBROISE-RENDU, A.-C. 1997. *Les faits divers dans la presse française de la fin du XIXe siècle. Étude de la mise en récit d'une réalité quotidienne.(1870-1910)*, Thèse de doctorat, Paris 1 Sorbonne, 765 p.

AUCLAIR, G. 1970. *Le Mana quotidien : structures et fonctions de la chronique des faits divers*, Paris, Anthropos, [réed. 1982, augmentée d'une préface de l'auteur et de l'essai *Le double imaginaire de la modernité dans la vie quotidienne*], 300 p.

BARTHES, R. 1964. Structure du fait divers, *Essais critiques*, Paris, Seuil, pp. 188-197.

KALIFA, D. 1995. *L'encre et le sang : récits de crimes et société à la Belle Époque*, Paris, Fayard, 351 p.

KAYSER, J. 1967. *Le quotidien français*, Paris, Armand Colin, coll. "Cahiers de la fondation nationale des sciences politiques", 169 p.

M'SILI, M. 2000. *Le fait divers en république*, Paris, CNRS Éditions, 311 p.

DES GLOSES DE MOT AUX TYPES DE TEXTES : UN BILAN DIFFERENCIÉ

Augusta MELA et Mathieu ROCHE
Université Montpellier III/LIRMM, Université Montpellier II/LIRMM

SOMMAIRE

Résumé : *Qu'il s'agisse d'évoquer une langue étrangère ou spécialisée (1), de procéder à une explication didactique (2), ou de s'assurer que l'interlocuteur attribue la signification adéquate au mot employé, le langage courant fournit de nombreux exemples de commentaires en situation parenthétique qui traduisent, expliquent le sens des mots en discours :*

(1) Si l'on admet que l'état d'un électron n'est pas entièrement décrit par sa position et sa vitesse de translation dans l'espace, mais qu'il est animé en outre d'un pivotement sur lui-même, ou « spin »*, mouvement essentiellement quantifié : son moment cinétique est d'un demi quantum et crée un moment magnétique égal à un magnéton de Bohr. (Hist. gen. sciences, 1964, t.3, vol.2)*

(2) Ce sont la dépigmentation, c'est-à-dire l'absence quasi totale des éléments colorés dermiques qui s'opposent normalement à l'action nocive des rayons ultra-violets d'origine solaire et l'anophtalmie, ou réduction de l'appareil oculaire allant le plus souvent jusqu'à sa disparition complète*. (Geze, La spéléologie scientifique, 1965)*

On appelle gloses *ces commentaires sur le mot, qui nous mènent au sens par des « chemins buissonniers » (Steuckardt et Niklas-Salminen,2005) .*

Les gloses se manifestent dans les textes par des marques typographiques comme les guillemets et des éléments lexicaux comme appelé, c'est-à-dire, ou*. Ces marqueurs sont polysémiques mais en tirant parti des particularités syntaxico-sémantiques des constructions segment glosé-marqueur-glose, leur repérage automatique est réalisable.*

De plus, ces marqueurs explicitent la nature du lien sémantique entre le segment glosé et sa glose : équivalence avec c'est-à-dire*,* ou *; spécification du sens avec* au sens *; nomination avec* dit*,* alias*,* baptisé *; hyponymie avec* en particulier, comme, tel, par exemple *; hyperonymie avec* et/ou autre(s)*.*

De premières études ont été réalisées sur la base Frantext (Mela,2004 ; Mela,2005). Nous élargissons ici nos investigations à d'autres types de textes, d'autres marqueurs et d'autres liens sémantiques. Des comparaisons avec la langue anglaise sont menées.

Nous répondons aux questions suivantes :

1) dans quelle mesure peut-on repérer automatiquement les gloses ?

2) dans quelle mesure tel marqueur de glose déclare tel lien sémantique entre le segment glosé et sa glose ?

3) la fréquence de gloses dépend du type de textes (poésie versus *ouvrage didactique) et pour un même lien sémantique, le choix du marqueur dépend du niveau de langue et du type de texte (ouvrage didactique* versus *roman) : dans quelles mesures ?*

en dressant un bilan, par types de textes, par marqueurs et par liens, de ces mesures. Une démonstration du logiciel de repérage sera proposée.

1. Introduction

Le travail que nous présentons a pour origine un questionnement de collègues linguistes : dans le cadre d'une recherche sur « le mot et sa glose », elles se demandaient si le repérage automatique des gloses de mot était possible. Les gloses de mot sont ces commentaires, souvent introduits par des termes de liaison tels que *c'est-à-dire*, *ou*, *autrement dit, ce qu'on appelle*, etc. qui définissent des mots en discours, comme la séquence en italique de l'énoncé suivant :

> (1)« 10 % de ces embauches vont porter sur un métier qui monte : le « testing », *c'est-à-dire la maîtrise des méthodologies rigoureuses de test des logiciels* », indique Dominique Duflo, le DRH. (L'Expansion, Avril 2006, p.136)

Un repérage automatique permettrait de ramener plus efficacement les gloses des corpus et de quantifier les phénomènes observés. De plus, dans la perspective d'un travail lexicographique, on espère, étant donné un mot spécifique, pouvoir repérer ses gloses et accéder ainsi à son sens. En effet, si on considère les énoncés suivants :

> (2)Quant aux espèces endobiontes des substrats meubles, *appelées* fouisseuses, elles sont légion, ce qui est normal car il est évidemment plus facile de fouir un sable ou une vase que de forer une roche, même tendre. (Peres.J-M, La vie dans l'océan,1966, p.72)

> (3)Les fouisseuses *sont* des espèces endobiontes des substrats meubles. (exemple construit)

on constate que l'apposition de la glose *appelées fouisseuses* contient, au même titre que la copule *sont* dans l'énoncé définitoire (3), une définition dont *fouisseuses* est le défini (definiendum) et *espèces endobiontes des substrats meubles*, le définissant (definiens).

Ainsi, les gloses de mot sont utiles, au même titre que les définitions, dans l'aide à l'acquisition terminologique, et ce à double titre :

- elles pointent le vocabulaire spécifique et les nouvelles unités qui demandent à être expliquées ;
- elles signent/signalent un texte définitoire comme en (4), voire elles définissent elles-mêmes le mot (1,2) :

> (4)Chaînon manquant entre l'apparition de la photographie et le cinéma des Lumière en 1895 , le flipbook (de flip over, feuilleter), aussi *appelé* folioscope, est ce livre animé, dont l'assemblage d'images dans le défilement donne l'impression du mouvement. (Libération, Stéphanie Binet, 19 mai 2006)

Nous pensons que leur repérage automatique est plus aisé que celui d'autres énoncés définitoires parce que les pivots du repérage, à savoir les marqueurs *appelé, c'est-à-dire,* etc. sont plus filtrants que la copule *être*, par exemple ; et parce que leur configuration, en apposition au mot glosé, est moins sujette à la variation que ne peuvent l'être les configurations où la définition est en prédication principale.

Nous proposons ici de faire le point sur ces questions de traitement automatique, en mettant les attentes en regard des outillages nécessaires.

Après avoir situé les gloses dans le champ des énoncés définitoires (§ 2), nous illustrons notre démarche en prenant pour fil conducteur l'exemple des indications de dénomination introduites par le marqueur *appelé.* Partant de la description linguistique de cette configuration (§ 3.1), nous enchaînons sur la modélisation et l'implémentation de son repérage (§ 3.2), nous évaluons et analysons les résultats obtenus (§ 3.3).

Nous serons alors à même de répondre à nos collègues linguistes : le repérage des gloses est possible ; la recherche des gloses d'un mot spécifique, comme l'extraction des définitions, est également possible, mais nécessite une analyse syntaxique partielle robuste préalable pour être traitée proprement.

2. Les gloses de mot dans le champ des énoncés définitoires

Menée à l'Université de Provence au cours des quatre dernières années, les études de la glose ont donné lieu à la publication de deux ouvrages collectifs [Steuckardt et Niklas-Salminen, 2003] et [Steuckardt et Niklas-Salminen, 2005]. Selon Agnès Steuckardt, directrice du projet :

> « Pour éclairer le sens d'un mot, l'analyse de corpus privilégie traditionnellement le repérage des associations récurrentes, sans se soucier de la conscience métalinguistique qu'en a ou non

> le locuteur. À la lumière de l'expérience concrète du travail sur les concordances de mot, il nous a semblé possible de trouver dans les gloses données par les locuteurs aux mots qu'ils emploient un autre accès au sens lexical. »

Une synthèse en ligne [Steuckardt, 2006] présente les différentes configurations syntaxiques des gloses de mot et une typologie des marqueurs de glose.

2.1. Forme des gloses

Les gloses de mot sont des configurations définitoires non formelles, parenthétiques. Sur le continuum définitionnel proposé par J. Rebeyrolle [Rebeyrolle, 1990, p.89] qui va des définitions directes aux définitions indirectes [Riegel, 1987], les gloses à marqueurs lexicaux (*appelé*, *c'est-à-dire*, *ou*, etc.) se situent au centre, entre les définitions directes (5,6) et les gloses à marqueurs typographiques (7) tels que virgule, parenthèses, crochets, deux points, tiret, etc.

> (5)*Nous appelons* donc définition (D) la mise en relation d'un terme à définir (A) et d'une séquence qui en est la paraphrase, constituée d'un second terme (B) auquel s'adjoint un ensemble de propriétés distinctives (X). (Rebeyrolle, 2000, p.89)
>
> (6)Avec ironie et non sans excès, *on appelle* « Khmers verts » les groupes d'écologistes qui ont contribué à substituer au corps des ingénieurs et urbanistes qui bitumaient et bétonnaient, un corps d'ingénieurs et d'urbanistes qui découpent avec autant de zèle les chaussées en zones d'exclusion (voitures, piétons, autobus, vélos), multiplient terre-pleins et espaces de verdure... Tous dispositifs qui, paradoxalement, affectent les qualités de partage et de rencontre des espaces publics. (Le Monde, Frédéric Edelmann, 27 avril 2006)
>
> (7)Depuis qu'il a été forcé par la Cour suprême des Etats-Unis de reconnaître la victoire de George W. Bush à l'élection présidentielle de 2000, Al Gore se consacre à la présentation d'une conférence illustrée (il l'appelle son « slide show *», sa soirée diapo*) démontrant la réalité du réchauffement global de la planète et l'urgence qu'il y a à le corriger. (Le Monde, Thomas Sotinel, 23 mai 2006)

Alors que la définition est la prédication principale des définitions directes (3,5,6), dans les gloses (1,2), la définition est en prédication seconde. Ce sont des définitions comme accidentelles, parenthétiques, insérées « en passant ». Le concept de glose rejoint le concept de *définition non formelle* dite *de substitution*, défini par J. Flowerdew [Flowerdew, 1992, 101] :

> « In contrast to formal and semi-formal definitions, *substitutions* are used most commonly where the definition is not the main focus of the discourse. Instead, they occur "embedded" in the overall discourse structure, inserted, so to speak, "en passant"; their function is not to provide important new information as such, but is a metalinguistic one to explain terms that arise as the lecture progresses. In this way they act as a sort of lubricating device which facilitates comprehension on the part of hearers, as they negotiate their way through the discourse. In the following example, for instance, definitions of the terms "anterior" and "posterior" are embedded in a more general description of a biological specimen:
>
> ... this is the anterior end/ anterior meaning front/ and posterior meaning behind ... »

2.2. Fonctions définitoires des gloses de mot

Si on analyse les énoncés suivants :

> (2)Quant aux espèces endobiontes des substrats meubles, *appelées* fouisseuses, elles sont légion, ce qui est normal car…
>
> (8)L'Arbadetorne, *qui signifie* l'herbe à détourne, a le pouvoir de détourner de son chemin celui qui marche dessus. (Sud Ouest, Karine Robin, 26 mai 2006)
>
> (8)Pour être précis, Manuel Desdín, Cubain de 21 ans, est "atlichnik", *autrement dit* "champion" à l'Institut de physique des basses températures de Kharkov. (Le Monde, Jésus Díaz, 26 mai 2006)

on constate que les gloses, comme les définitions directes, peuvent être :

- explicatives : c'est le cas dans les indications de signifié (8,1) et de nouvelle nomination (9), qui amènent le récepteur de l'inconnu vers le connu ;
- didactique : c'est le cas dans les indications de dénomination (2,4) qui amènent le récepteur du connu vers l'inconnu.

Comme nous le fait observer A. Steuckardt [Steuckardt, 2006] :

> « Visée didactique et visée explicative sont des cheminements inverses d'un même parcours. Du point de vue de l'analyste, l'un comme l'autre ouvrent un accès au sens lexical. »

3. Le traitement informatique

Nous nous intéresserons ici au cas des gloses dites, par raccourci, à « marqueurs » lexicaux. Ces termes de liaison *c'est-à-dire*, *à savoir*, *ou*, a*ppelé*, etc. sont polysémiques mais alliés aux propriétés syntaxico-sémantiques des configurations où ils interviennent en tant qu'introducteurs de glose de mot, ils pourront « marquer » les gloses. Ainsi, lorsqu'il est en configuration de glose comme dans l'énoncé qui suit :

> (9)De remarquables travaux, qui n'avaient pas trouvé d' écho à leur parution, au cours du xixe siècle, ont pris toute leur signification à l' aurore du XXe et ont formé les bases de toute une discipline nouvelle, qui a pris une ampleur et une importance considérables, la génétique, *ou science de l'hérédité*. (Anonyme, Hist. gen. sciences, t.3 vol.1, 1961, p.550, Frantext)

le terme *ou* joint un terme en usage (*génétique* ici) à un terme en mention (*science de l'hérédité)* ; le terme en mention s'applique métalinguistiquement au premier ; il est co-possible et son statut métalinguistique est marqué par une absence de détermination [Tamba, 1987, p.27-28]. Projeté sur le corpus de *l'Histoire générale des scien*ces de Frantext[1] [Mela, 2004], le patron « *ou* précédé d'une ponctuation et suivi d'un substantif non déterminé » permet de ramener des gloses en *ou* telles que (10) avec une précision[2] de 97%.

[Mela, 2005] traite des gloses en *dit* dans l'environnement Frantext-Stella. Nous détaillons ici le cas des indications de dénomination introduites par *appelé*, sur des corpus diversifiés.

Notons que nous procédons actuellement marqueur par marqueur mais que ces marqueurs pourront agir de concert pour détecter tous types de gloses en une seule passe sur le corpus.

Outre le repérage d'un (ou plusieurs) types de gloses, d'autres fonctionnalités seraient utiles :

- la repérage des gloses d'un mot donné ;
- l'extraction des arguments de la définition, le défini (definiendum) et le définissant (definiens).

Ces trois fonctionnalités ne comportent pas les mêmes difficultés pour des raisons que nous analysons en 5.3. Actuellement seule la première fonctionnalité est implémentée. Nous en évaluons les résultats ; nous spécifions les deux autres fonctionnalités.

Notre article suit les étapes suivantes :

- en section 2, nous définissons le concept de glose ;
- en section 3, nous partons de la description linguistique de la configuration recherchée, pour aboutir au patron, ou motif de recherche ;
- en section 4, le patron est implanté et projeté sur des corpus annotés morpho-syntaxiquement. Nous analysons les résultats.

3.1. Description linguistique de *appelé* dans la glose de dénomination

Le verbe *appeler* appartient à la table 11 de Gross [Gross, 1975]. Les verbes de cette table ont un complément direct substantival et un complément indirect en *à* (*appeler à voter/aux urnes/à ce que Phrase*). Dans la construction indirecte, *appeler* ne dénomme pas. La dénomination peut être en prédication première dans la configuration N_0[3] *appeler* N_1 N_2, ou en prédication seconde dans la configuration N_1, *appelé* N_2. Le participe passé *appelé* peut apparaître dans une autre configuration parenthétique *X, (ce) qu'on a appelé Y*. Mais le plus souvent, il s'agit de simples modalisations autonymiques[4]. Nous avons écarté cette configuration pour l'instant.

Dans la configuration N_1 *appelé* N_2 , *appelé* est en position d'épithète détaché (2) ou pas selon qu'il est séparé ou non du mot glosé par une virgule, dont il est peut être séparé par un adverbe

---

[1] La base Frantext est accessible par abonnement à l'adresse :<http://www.frantext.fr/>.

[2] Le terme précision est défini en section 4.

[3] N_0 : sujet formel, N_1 et N_2 sont les compléments du verbe, leur ordre correspondant à leur propriété de présence : obligatoire à facultative (cf. [Gross, 1975, p.13]).

[4] L'*autonymie* peut se définir comme la « propriété linguistique en vertu de laquelle tout mot ou tout élément linguistique peut être employé pour se désigner lui-même ». Par exemple : dans « rose » a 4 lettres, *rose* est autonyme. On parle du signe *rose* et non de ce qu'il dénote, une fleur. La *modalisation autonymique* est un cas particulier de l'autonymie. J. Authier-Revuz [Authier-Revuz, 1995] la définit comme « une opacification, résultant de ou consistant en – selon que l'on parle du résultat ou du processus énonciatif – une référence au monde accomplie en interposant sur le 'trajet' de la nomination la considération de l'objet signe par lequel on réfère. » Par exemple : dans *Le risque existe également de ce qu'on a appelé la «malédiction pétrolière»* (Libération**,** Rueff Judith**,** 17 mai 2006), les guillemets font que l'on s'arrête sur le mot et que le terme entre guillemets n'est plus complètement « transparent », il y a « opacification ».

Dans la base Frantext entière, sur 11occurrences de « ,(*ce) qu'on a appelé* », deux sont des gloses de mot, une est une reformulation de proposition et les autres, des modalisations autonymiques.

comme en (4) : *le flipbook, aussi appelé folioscope*. Il ne se compose pas avec les auxiliaires *être* et *avoir* (*j'ai appelé, est appelé*).
Le mot glosé N_1 est recteur : il impose les contraintes d'accord en genre et nombre. Dans des cas plus rares et à condition que le mot glosé N_1 soit sujet, *appelé* peut lui être antéposé :

> (10)Appelée « bancor », l'unité de compte proposée pour comptabiliser les créances et les dettes était totalement fiduciaire malgré le préfixe (sic) « or ». (1960, *L'univers économique et social*, François Perroux éd., Frantext) (cité dans [Steuckardt, 2006])

Dans tous les cas, N_2 est attribut du terme recteur N_1.
Enfin, il attend un objet direct (c'est-à-dire sans préposition, ni déterminant contracté).

3.2. De la description linguistique à l'implémentation

Notre repérage de la glose de dénomination en *appelé* s'appuie sur un étiquetage morpho-syntaxique des mots du corpus. Puisque nous ne disposons pas de structuration syntagmatique du texte, le motif de filtrage doit donc coller à la linéarité du texte.
Le schéma abstrait *X(,)appelé Y* n'est pas opérationnel pour plusieurs raisons :

- on a vu qu'il ne couvrait pas les cas d'antéposition de *appelé* (11) ;
- par ailleurs, il suppose que X et Y sont des groupes substantivaux ;
- des éléments peuvent s'insérer entre le pivot verbal *appelé* et X et Y : adverbe, incise avant (4) ou après (13) *appelé,* coordination (14) :

> (4)Chaînon manquant entre l'apparition de la photographie et le cinéma des Lumière en 1895 , le flipbook (de flip over, feuilleter), aussi *appelé* folioscope, est ce livre animé, dont l'assemblage d' images dans le défilement donne l'impression du mouvement. (Libération, Stéphanie Binet, 19 mai 2006)
>
> (13) L' idée est la suivante : à l'espace des hypothèses est associé un autre espace *appelé*, pour des raisons de similarité avec les mécanismes de l'évolution naturelle, espace "génotypique". (Corpus Scientifique)
>
> (11) Avec ce "radio-conducteur", perfectionné en 1890 et *appelé* «cohéreur» par Lodge, la radioélectricité était née. (<http://www.elec.unice.fr/pages/phototheque/photos.html>)

La présence de la virgule devant *appelé* ne nous a pas parue significative. Pour toutes ces raisons, nous nous en sommes tenus au simple patron :

```
(P)  appelé(e)(s)/Participe_passé suivi d'un mot autre que (à/au/aux)
```

Nos corpus sont étiquetés avec l'étiqueteur WinBrill[1] [Brill, 1994]. Aux accords en genre et nombre près, WinBrill distingue 5 sortes de participes passés :

EPAR : (sg|pl)	Verbe « être », non conjugué, participe passé
APAR : (sg|pl)	Verbe « avoir », non conjugué, participe passé
VPAR : (sg|pl)	autre Verbe, non conjugué, participe passé après « avoir »
ADJ1PAR : (sg|pl)	Participe passé après « être », adjectival ou verbal
ADJ2PAR : (sg|pl)	Participe passé adjectival, singulier (non après auxiliaire)

Pour *appelé*, trois étiquettes sont donc disponibles : VPAR:(sg|pl), ADJ1PAR:(sg|pl) et ADJ2PAR:(sg|pl). Dans la configuration qui nous intéresse, *appelé* n'est pas composé. Il sera donc, sauf erreur d'étiquetage, étiqueté ADJ2PAR.
Traduit en langage des expressions régulières Perl, notre patron (P) devient :

```
(P)_version Perl    appelée?s? \/ADJ2PAR:(sg|pl) (?!(à|au|aux)
```

4. Expérimentations

4.1. Corpus utilisés

Nous avons utilisé trois corpus de domaines différents :

- *Corpus Scientifique :* le livre « Apprentissage Artificiel[2]» d'Antoine Cornuéjols et Laurent Miclet (éditions Eyrolles) sur lequel nous nous appuyons dans nos expérimentations est un corpus utilisé et prétraité dans le cadre de DEFT'06[3], défi francophone de fouille de textes ;

[1] On peut en savoir plus sur WinBrill et le télécharger gratuitement à partir du site de l'ATILF (http://www.atilf.fr).
[2] http://www.editions-eyrolles.com/Livre/9782212110203.
[3] DEfi Fouille de Textes : <http://www.lri.fr/ia/fdt/DEFT06/>.

- *Corpus Journalistique :* 71 articles de presse du 20 mai 2006 relevés sur Europresse[1] ;
- *Corpus Littéraire :* un sous-corpus de Frantext réduit à l'année 1950, tous genres sauf théâtre et poésie (14 œuvres) ;

à partir desquels, trois traitements sont effectués :

- Un premier filtre morphologique sélectionne les phrases contenant des occurrences de « appelé (e) (s) » ;
- L'étiqueteur de Brill est appliqué sur ces phrases préalablement mises au format ;
- Le patron morpho-syntaxique (P) est appliqué et départage deux listes de phrases. Les phrases reconnues par le filtre, sont appelées Positifs : elles sont censées contenir des gloses de dénomination. Les phrases non retenues, les Négatifs, sont censées ne pas contenir de gloses de dénomination ;
- Une évaluation peut alors être effectuée sur chacune des deux listes :
    - Elle consiste à vérifier si les Positifs sont réellement des gloses de dénomination ou pas. Lorsque ce sont des gloses de dénomination, ce sont des Vrais Positifs, notés *VP*, et nous les classons en 3 sous-groupes : définitions, synonymies, hyperonymies. Lorsque ce ne sont pas des dénominations, ce sont des Faux Positifs, notés *FP*. Nous les classons également en sous-groupes, suivant qu'il s'agit de dénominations propres par NPs et autres désignateurs rigides ou que les sources de mauvais aiguillage soient autres : étiquetage erroné, présence d'un complément introduit par une préposition autre que *à*, etc.
    - Nous examinons également les Négatifs, pour vérifier que tous le sont vraiment et sinon analyser les raisons du « silence ». Les Faux Négatifs, notés *FN*, sont alors les gloses de dénomination passées sous silence. Les autres sont les Vrais Négatifs, notés *VN*.

4.2. Mesures d'évaluation

La mesure d'évaluation appelée *précision* doit être calculée. Une telle mesure fréquemment appliquée dans le domaine de l'apprentissage artificiel [Cornuéjols et Miclet, 2002] est donnée par la formule ci-dessous :

$$précision = VP/(VP+FP)$$

Une précision de 100% signifierait que toutes les phrases ramenées par notre patron sont des gloses de dénomination. Si elle ne vaut pas 100%, c'est qu'il y a du « bruit ».

Pour vérifier que notre patron ramène toutes les gloses de dénomination du corpus, nous utilisons une autre mesure d'évaluation appelée *rappel* [Cornuéjols et Miclet, 2002]. Pour ce faire, une évaluation « manuelle » consiste à examiner cette fois les Négatifs et à vérifier que tous le sont vraiment et sinon analyser les raisons du « silence ». Les Faux Négatifs seraient les gloses de dénomination passées sous silence.

Le rappel est donné par la formule ci-dessous :

$$rappel = VP/(VP+FN)$$

Un rappel de 100% signifierait que toutes les gloses de dénomination ont été extraites du corpus.

Notons que la somme des positifs et négatifs (*nb =VP+FP+VN+FN*) recouvre l'ensemble du corpus à filtrer.

En règle générale, il est important de déterminer un compromis entre le rappel et la précision. Pour cela, nous pouvons utiliser une mesure prenant en compte ces deux critères d'évaluation en calculant le *Fscore* [Rijsbergen, 1979] :

$$Fscore = (\beta^2+1)*précision*rappel/(\beta^2*précision+rappel)$$

Le paramètre β de la formule ci-dessus permet de régler les influences respectives de la précision et du Rappel. Il est très souvent fixé à 1 pour accorder le même poids à ces deux mesures d'évaluation.

4.3. Évaluation de (P) sur les trois corpus

4.3.1. Prélimaire : Où s'arrêtent les gloses de dénomination ?

La délimitation précise des gloses recherchées est guidée par la tâche à réaliser. On cherche des gloses pour trouver du sens. Si c'est du sens lexical, nous nous intéresserons davantage à la

[1] Europresse, <http://www.bpe.europresse.com/>, correspond à 19 titres de la presse nationale et régionale, dont Le Monde, Les Echos, Libération, Sud-Ouest, etc.

fonction *catégorisante* qu'à la fonction *individualisante*[1] des noms, donc nous privilégierons les dénominations communes alors que les désignations rigides, par noms propres telles que *un psychologue américain appelé Gesell*, ou autres noms de lieu, d'enseigne(15), de produit etc. seront de moindre intérêt. En revanche, si l'on cherchait à constituer un dictionnaire de noms propres, elles deviendraient intéressantes.

> (12) Le comité de salut public arrête que la maison nationale ci-devant *appelée les menus-plaisirs*, située rue Bergère, servira désormais pour l' institut national de musique établi par les décrets de la convention nationale. (Anonyme,Enseign. mus. 1. enseign. off., 1950 p.8, Frantext)

De plus, le sens trouvé doit être le plus riche possible. La définition, la plus complète possible. Telle qu'elle est définie en (5), la définition correspond au schéma A=B+C, où A est le terme défini, B une classe (l'hyperonyme) et C la caractéristique, la « détermination distinctive de B[2] ». Ce schéma est illustré par l'énoncé suivant :

> (13) On appelle carré un losange dont les côtés sont égaux et les angles sont droits.
>
> où A=*carré*, B=*losange* et C = *dont les côtés sont égaux et les angles sont droits*.

Face à l'énoncé suivant :

> (14) Il utilise pour ce faire un mécanisme *appelé* chunking qui produit des "chunks" (littéralement "gros morceaux") ou macro-opérateurs. (Corpus Scientifique)

si on prend le schéma de glose *X marqueur Y* au pied de la lettre, X = *un mécanisme* et Y =*chunking*, le schéma correspond[3] à une définition pauvre : *le chunking est un mécanisme.* On perd l'information donnée par la détermination distinctive C.
Si au contraire, on considère *X marqueur Y* comme un schéma abstrait, qui ne colle pas nécessairement à la linéarité du texte, alors, appliqué à l'énoncé (17), la correspondance entre schéma de glose et schéma de définition est totale : *B+C*↔*X= un mécanisme qui produit des chunks* … et A↔Y= *chunking*. On rend compte de la définition complète contenue dans (17).
Autrement dit, on doit admettre que dans sa réalisation textuelle linéaire, le groupe substantival X du schéma *X marqueur Y* peut être discontinu, sa tête restant à gauche du marqueur et recteur de celui-ci, et ses compléments (ppassé, pprésent, relative) étant placés à droite de la glose.
Pour ces raisons, nous avons considéré que les cas tels que (17) étaient des définitions prototypiques (étiquetées *D* dans notre évaluation[4]).
Notons qu'en abstrayant le schéma de glose comme nous venons de l'argumenter, nous continuons au niveau de l'analyse sémantique, la démarche que nous avons dû entamer lors de l'analyse syntaxique (§ 3.2). Mais ce faisant, nous faisons un écart : telle qu'elle est définie actuellement dans le projet aixois, la glose n'englobe pas les cas d'hyper/hyponymie. Ainsi, pour A. Steuckardt[5], la séquence *le textile connu sous le nom de ramie* de l'énoncé suivant :

> (15) Le textile *connu sous le nom de ramie* provient de l'écorce des tiges de l'ortie de Chine, *appelée encore ramie* (Blanquet, Technol. mét. habill., 1948, p. 51, in TLFI, s.v. *RAMIE*).

n'est pas une glose parce que *ramie* n'est pas une dénomination de *textile*, les deux termes *ramie* et *textile* étant simplement dans un rapport hyperonyme-hyponyme. Pour nous, cette séquence est intéressante pour l'acquisition lexicographique, parce qu'elle contient la définition *la ramie est un textile qui provient de l'écorce des tiges de l'ortie de Chine (appelée encore ramie)*.

4.3.2. Évaluation du patron (P)

Le patron (P) que nous évaluons résulte de la phase « manuelle » d'analyse linguistique. Étant établi uniquement sur des critères linguistiques généraux, il est lâche et son rappel est maximal (100%, cf. tableau ci-dessous). Sauf dans le cas du corpus Littéraire où il chute de 25% à cause d'un seul cas de silence, dû à un mauvais étiquetage, mais le nombre de gloses de dénomination est trop faible sur ce corpus pour que le résultat soit significatif.
Des allers et retours entre analyse linguistique et résultats sur corpus permettront d'améliorer la précision du repérage et les scores obtenus, tout en gardant un rappel acceptable.

---

[1] Ces termes sont de P.Siblot et cités dans [Kleiber, 1996, p.584].
[2] Le terme est de Kleiber et Tamba [Kleiber et Tamba, 1990, p. 24].
[3] Nous utilisons les symboles « ↔ » pour la correspondance entre le schéma de la définition et le schéma de glose ; « = » pour la correspondance entre le schéma de glose et sa réalisation textuelle.
[4] Les corpus et le détail des évaluations sont disponibles sur :
<http://www.lirmm.fr/~mroche/Recherche/glose.html>
[5] Communication personnelle (avril 2006).

Le patron (P) est un point de départ et constitue un dispositif dont on peut faire varier les paramètres (précision et rappel) en fonction de la tâche poursuivie et de la taille du corpus examiné : une analyse linguistique manuelle sur un petit corpus privilégiera un rappel maximal ; alors que l'extraction automatique de définitions à partir d'un gros corpus aura intérêt à resserrer le patron, quitte à laisser sous silence une partie des gloses hors calibre. Nous détaillons en §5 comment le dispositif peut être adapté.
Le tableau ci-dessous présente le résultat global obtenu pour (P) à partir des trois corpus. Ce tableau montre que les résultats les plus significatifs en terme de Fscore sont obtenus sur le corpus Scientifique. En effet, avec un tel corpus très spécialisé, il est nécessaire de définir des termes. Sur l'ensemble des corpus, le rappel est élevé montrant que le patron (P) ramène toutes les gloses pertinentes (moins une). La faible précision pour les corpus Journalistique et Littéraire est dûe à une présence importante de dénominations propres, que nous n'avons pas cherché à exclure dans un premier temps.

Corpus	***nb***	***VP***	***FP***	***VN***	***FN***	***Précision***	***Rappel***	***Fscore***
Scientifique	70	25	8	37	0	75,8%	100%	**86,2%**
Journalistique	51	6	10	35	0	37,5%	100%	**54,5%**
Littéraire	58	3	11	44	1	21,4%	75%	**33,5%**

5. Analyse des résultats

Nous analysons les résultats sous l'angle des améliorations du repérage, des types de gloses obtenus, de l'extraction du definiendum Y et du definiens X, et de la recherche des gloses d'un mot donné.

5.1. Améliorations possibles du repérage

Plusieurs pistes amélioreraient sensiblement les scores précédents. Certaines sont à notre portée, d'autres demandent une investigation plus poussée.
Le patron (P) actuel n'exclut que la préposition *à* et les déterminants contractés avec *à*. Nous ne souhaitions pas exclure a priori les énoncés où des incises débutant par une préposition s'insèrent entre *appelé* et son complément direct, comme dans l'énoncé suivant :

> (16)On y parvenait par un escalier en bois blanc *appelé*, dans l'argot du bâtiment, échelle de meunier. (Balzac.H de, Le cousin Pons, 1847, p.751, Frantext)

Cette stratégie s'avère coûteuse sur nos corpus puisque nous n'avons ramené aucun de ces cas alors que nous comptons 3 cas de bruit en présence de préposition tels que :

> (17)Les clubs sont fondés à souhaiter que des assurances soient prises pour leurs joueurs *appelés en* sélection et à réclamer une indemnisation. (L'Équipe, 19 mai 2006, Corpus Journalistique)

Il semble donc qu'on ait intérêt, quitte à laisser sous silence quelques cas, à étendre l'interdiction à toutes les prépositions.
Les deux propositions suivantes amélioreraient également la précision mais toujours au détriment du rappel, car elles nécessitent de prévoir une position pour le definiendum Y (cf. §5.3) dans le patron qui deviendrait par exemple :

```
appelé(e)(s) ?ADV ( ?! Prep) Y_Substantival
```

Les modalisations autonymiques (B-MA dans notre évaluation) telles que (21) pourraient être évitées en excluant les adjectifs de la position du definiendum Y :

> (18)J'entre dans le monde *appelé réel* comme on entre dans de la brume. (Green.J, Journal T.5, 1950, p. 267, Corpus Littéraire)

Les dénominations rigides, notées *DP* dans notre évaluation, constituent la cause principale (plus de 2/3 des FP) du bruit actuel. On pourrait, à condition qu'ils soient reconnus en amont, exclure les noms propres de la position Y. Une étude contrastive des dénominations communes versus rigides (présence ou pas du déterminant, etc.) reste à faire pour voir si un filtrage des dénominations rigides comme *les menus-plaisirs* de (15) est faisable.

5.2. Quels types de glose obtient-on ?

Parmi les Vrais Positifs, outre les définitions complètes (17), on obtient de simples synonymies (étiquetées *S* dans notre évaluation) comme celle qui lie *praticien* et *recors*[1], dans l'énoncé suivant :

> (19)Le praticien, vulgairement *appelé* recors est l'homme de justice par hasard, il est là pour assister l'exécution des jugements, c'est, pour les affaires civiles, un bourreau d'occasion. (Balzac, Cous. Pons, 1847, p. 173, Corpus Littérraire)

Les cas d'hyperonymie/hyponymie (étiquetés *H* dans notre évaluation) sont ceux où un hyperonyme de Y précède *appelé* et où la détermination distinctive est nulle :

> (20)Nous ne respirons peut-être pas à votre hauteur, mais nous avons un viscère *appelé* cœur. (Bazin H., La mort du petit cheval, 1950, p.106, XIII, Corpus Littéraire)

Souvent la glose se greffe sur une autre glose ou sur un énoncé définitoire. Dans ce cas, deux niveaux de définition co-opèrent. Ainsi, dans l'énoncé (22), par un jeu de circulation sémantique, le definiens de la prédication principale rejoint le definiens de la prédication seconde, les deux definiens s'amplifiant mutuellement.

5.3. L'extraction du definiens X et du definiendum Y est-elle possible ?

L'extraction du definiens X et du definiendum Y nécessite de prévoir les positions de X et à Y dans le patron. Si ces positions sont contigües au pivot verbal, le patron sera trop strict. Si on autorise des fenêtres entre *appelé*, X et Y – douze mots pour rendre compte de (13), le patron risque d'être trop lâche.

Par ailleurs, du fait que X est souvent discontinu (cas de fausses hyper/hyponymies tels que (17)), la question du rattachement de la détermination distinctive se pose.

Pour être traitées proprement, ces deux tâches nécessitent une analyse structurelle des corpus, au moins partielle. Plutôt que de raisonner en termes de fenêtre de mots, on pourrait alors raisonner en termes de présence optionnelle d'un groupe syntagmatique. La question du rattachement de la détermination distinctive de X reviendrait à reconnaître juste après *appelé* les configurations possibles de cette détermination distinctive : proposition relative, proposition construite autour d'un participe passé ou d'un gérondif.

Autre problème, lié cette fois à l'ambiguïté structurelle de la langue : comment reconnaître automatiquement l'empan du terme glosé lorsque des GNs enchassés précèdent *appelé* (24,25) ? On pourrait s'appuyer sur des marques d'accord grammatical, ou typographiques comme les guillemets en (24), mais ces marques ne sont pas toujours disponibles (25) :

> (21)Le jour du drame ces deux employés étaient occupés à des travaux de maintenance sur les pompes du réseau de captage des « lixiviats », autrement *appelés* « jus de décharge ». Ils avaient été retrouvés asphyxiés par les gaz délétères, sur une plate-forme de sécurité, installée dans un puits.(Sud Ouest, Axelle Maquin-Roy,19 mai 2006)
>
> (22)D'autant plus que les intermédiaires spécialisés dans la valorisation de sites (*appelés* "brownfields developers" aux États-Unis) sont plutôt rares en France, avec quelques exceptions, comme Terra Verde Capital... (La Tribune, 19 mai 2006)
>
> La question de la recherche des gloses d'un mot spécifique se ramène à la question 5.3. Ainsi, chercher la définition de *espace génotypique* de l'exemple (13) nécessite de pouvoir exprimer dans le patron de recherche que Y = *espace génotypique* ; pour cela il faut que Y (et sa position) soit prévue dans le patron.

5.3.1. Énoncés définitoires et typologie des textes

On sait ([Rebeyrolle, 2000, Chap8]) que l'on peut caractériser les textes à partir de leurs propriétés définitoires : la fréquence mais aussi la forme (prédication principale versus seconde) des définitions sont des indicateurs de types de corpus. Les textes didactiques sont propices aux définitions en prédication principale. Les définitions en prédication seconde y sont également nombreuses. À l'opposé, les textes poétiques ne contiennent ni définitions ni gloses. Dans Europresse, les définitions formelles en *appeler (On appelle X, X s'appelle*) sont rares. La plupart des définitions en *appeler* sont des gloses. La glose se rencontre aussi dans les textes littéraires de Frantext alors que la définition, posée en prédication première, y est moins représentée.

[1] recors : subst. masc. Terme du droit : Personne qui assistait un huissier dans les opérations d'exécution en qualité de témoin et dont la présence est aujourd'hui facultative. Synon. praticien.(TLFI)

On voit que, dans l'accession au sens, la voie par les gloses est complémentaire de la voie par la définition, tant par les configurations linguistiques utilisées que par le type de textes à exploiter.

6. Conclusion

Moyennant une adaptation à la tâche visée et au corpus traité, le patron linguistique (P) permet de repérer des gloses de dénomination. La recherche des gloses d'un mot spécifique, comme l'extraction des définitions, est également possible, mais nécessite une analyse syntaxique partielle robuste préalable pour être traitée proprement. Ce sera notre prochaine étape.

Les perspectives d'utilisation des énoncés définitoires sont multiples. Ils sont utiles pour l'acquisition lexicographique, terminologique. D'ores et déjà, on peut utiliser le Web comme dictionnaire encore plus efficacement en sélectionnant parmi les concordances d'un mot donné celles qui glosent le mot. La recherche de « webzine, c'est-à-dire » sur Google[1] place en 1ère position le résultat suivant :

> Résultats 1 - 8 sur un total d'environ 10 pour "webzine, c'est-à-dire". (0,23 secondes)
> Expose Libre - Webzine francophone dédié à Magic The Gathering
> C'est un **webzine, c'est à dire** un magazine normal, sauf qu'au lieu d'être imprimé et vendu dans des kiosques à journaux, celui ci est publié sur internet. **...**
> www.exposelibre.org/?numero=0&article=infos - 8k

alors que la recherche de « webzine » demande d'aller chercher la définition sur les sites référés. L'annotation linguistique du Web[2] [Kilgariff, 2003], [Kilgarriff et Grefensette, 2003] rendra la recherche de gloses de mot d'autant plus efficace.

L'étude des énoncés définitoires sur des corpus alignés [Pearson, 2000], [Suarez de la Torre, 2004] révèle que souvent les gloses à marqueur lexical sont traduites par des gloses sans marqueur lexical et vice-versa : grâce à l'alignement, le repérage des gloses d'un corpus peut alors servir à pointer, dans l'autre corpus, les définitions en correspondance, même si elles ne sont pas marquées lexicalement.

Des travaux pourraient être menés consistant à étudier les gloses présentes en corpus parallèles multilingues. En effet, le fait de détecter de manière efficace les gloses en français doit permettre l'observation de similitudes linguistiques selon les langues, étude particulièrement utile pour les applications de traduction automatique par exemple. Par ailleurs, une telle étude permettrait de mettre en oeuvre des méthodes de construction automatique de dictionnaires multilingues.

Nous remercions Agnès Steuckardt pour sa disponibilité et sa générosité intellectuelle, Antoine Cornuéjols et Laurent Miclet pour nous avoir confié la version électronique de leur ouvrage.

BIBLIOGRAPHIE

AUGER, A. 1997. *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles*, Thèse de doctorat, Université de Neuchâtel Disponible sur : <http://www.unige.ch/cyberdocuments/unine/theses2000/AugerA/these_body.html>. (Consulé le 28.05.2006)

AUTHIER-REVUZ, J. 1995. *Ces mots qui ne vont pas de soi*, Paris, Larousse.

BOUVERET, M. 1998. Approche de la dénomination en langue spécialisée, *Meta, XLIII,3*.

BRILL E. 1994. Some Advances in Transformation-Based Part of Speech Tagging, *Actes de AAAI, Vol. 1*, pp. 722-727. Disponible sur : http://citeseer.ist.psu.edu/brill94some.html

CORNUÉJOLS, A et MICLET, L. (avec la participation d'Yves KODRATOFF). 2002. *Apprentissage artificiel : Concepts et algorithmes*, Eyrolles.

FLOWERDEW, J.,1992. Salience in the performance of one speech act : the case of definitions, *Discourse Process*, 15 (2), pp. 165-181.

GROSS, M. , 1975. *Méthodes en syntaxe*, Paris, Hermann.

KILGARRIFF, A. et GREFENSETTE, G. 2003. Introduction to the Special Issue on the Web as Corpus, *Computational Linguistics*, vol. 29/3, pp. 333-347. <http://mitpress.mit.edu/>

[1] Test effectué le 4 Juin 2006.

[2] Cf.<http://citeseer.ist.psu.edu/568007.html> et WebCorp :<http://www.webcorp.org.uk/>

KILGARRIFF, A. 2003. Linguistic search engine, in Kiril-Simov, (éd.), Shallow Processing of Large Corpora : workshop tenu conjointement à *Corpus Linguistics 2003*, Lancaster, England, pp. 53-58. <http://citeseer.nj.nec.com/568007.html>
KOSKAS,E. et KREMIN, H. (resp.) 1984. La dénomination, *Langue française*, *76*.
KLEIBER, G., 1984. Dénomination et relations dénominatives, *Langages, 76*, La dénomination, Koskas et Kremin (resp.).
KLEIBER, G.et TAMBA, I. 1990. L'hyponymie revisitée : inclusion et hiérarchie, *Langue française*, *98*.
KLEIBER, G., 1996. Noms propres et noms communs :un problème de dénomination, *Meta,XLI,4*.
MELA, A. 2004. Linguistes et "talistes" peuvent coopérer : repérage et analyse des gloses, *Revue Française de Linguistique Appliquée, IX (1)*, *Linguistique et informatique : nouveaux défis*, B. Habert (resp.). Disponible sur <http://www.univ-montp3.fr/~amela/PUBLICATIONS/>
MELA, A. 2005. Les gloses de nomination seconde, *Les marqueurs de glose*, Aix-en-Provence, Publications de l'Université de Provence. Disponible sur :
<http://www.univ-montp3.fr/~amela/PUBLICATIONS/>
MORTUREUX, M-F. (resp.) 1990. L'hyponymie et l'hyperonymie, *Langue française*, 98.
PEARSON, J. 1999. Comment accéder aux éléments définitoires dans les textes spécialisés ?, *Terminologies nouvelles*, 19. Disponible sur : <http://www.rifal.org/3_information.html>. (Consulté le 25.05.2006)
PEARSON, J. 2000. Une tentative d'exploitation bi-directionnelle d'un corpus bilingue, *Cahiers de grammaire n°25*, Sémantique et corpus, A. Condamines (resp.). Disponible sur :
<http://www.univ-tlse2.fr/erss/>. (Consulté le 27.05.2006)
REBEYROLLE, J. 2000. *Forme et fonction de la définition en discours*, Thèse de Doctorat en Sciences du langage, Université de Toulouse-le-Mirail, Toulouse II. Disponible sur :
<http://www.univ-tlse2.fr/erss/>. (Consulté le 27.05.2006)
REBEYROLLE, J. et TANGUY, L. 2000. Repérage automatique de structures linguistiques en corpus : Le cas des énoncés définitoires, *Cahiers de grammaire n°25*, Sémantique et corpus, A. Condamines (resp.). Disponible sur :<http://www.univ-tlse2.fr/erss/>. (Consulté le 27.05.2006)
RIEGEL, M., PELLAT, J-C. et RIOUL, R. 1994. *Grammaire méthodique du français*, Paris, PUF.
RIEGEL, M. et TAMBA, I. (resp.) 1987. La reformulation du sens dans le discours, *Langue française*, 73.
RIEGEL, M. et TAMBA, I. 1987. Présentation, *Langue française*, 73, pp. 3-4.
RIEGEL, M. 1987. Définition directe et indirecte dans le langage ordinaire : les énoncés définitoires copulatifs, *Langue française*, 73, pp. 29-53.
STEUCKARDT, A. et NIKLAS-SALMINEN, A. 2003. *Le mot et sa glose*, Aix-en-Provence, Publications de l'Université de Provence.
STEUCKARDT, A. et NIKLAS-SALMINEN, A. 2005. *Les marqueurs de glose*, Aix-en-Provence, Publications de l'Université de Provence.
STEUCKARDT, A. 2006. Du discours au lexique : la glose, Séminaires de l'ATILF, Disponible sur : <http://www.atilf.fr/atilf/seminaires/historique.htm#Steuckardt_2006-03>. (Consulté le 27.05.2006)
SUAREZ DE LA TORRE, M.M. 2004. *Análisis contrastivo de la variación denominativa en textos especializados : del texto original al texto meta*. Tesis Doctoral, Universitat Pompeu Fabra. Disponible sur : <http://www.tdx.cbuc.es/TESIS_UPF/AVAILABLE/TDX-0217105-130025//tmst1de1.pdf>. (Consulté le 27.05.2006)
TAMBA, I. 1987. « Ou » dans les tours du type : « un bienfaiteur public ou évergète », *Langue française*, 73, pp. 16-28.
TAMBA-MECZ, I. 1994. *La sémantique*, Paris, PUF.
VAN-RISBERGEN, C.J. 1979. *Information Retrieval*, 2nd edition, London, Butterworths.

CONSTITUER ET EXPLOITER UN GRAND CORPUS ORAL : CHOIX ET ENJEUX THÉORIQUES. LE CAS DES ESLO[1]

Lotfi ABOUDA et Olivier BAUDE
CORAL – Université d'Orléans

SOMMAIRE

0. Introduction

Contrairement aux corpus de français écrit il n'existe pas de grand corpus de français oral disponible pour l'ensemble de la communauté scientifique. La présentation du projet de diffusion des corpus ESLO (Enquêtes Socio-Linguistique à Orléans) à l'ensemble des acteurs de la recherche, qu'ils viennent des sciences cognitives ou de l'anthropologie, de la physique (traitement du signal) ou des études de genre (gender studies), de la dictionnairique ou du TAL, est l'occasion d'interroger les raisons complexes d'une telle situation.

Un regard épistémologique, notamment sur la place des données en linguistique, apporte des éléments d'explication qu'il convient de prendre en compte avant de proposer des pistes pour une méthodologie favorable à l'exploitation de grands corpus oraux. Les principaux choix théoriques et techniques opérés lors de l'exploitation scientifique (numérisation, transcription, annotation, diffusion, analyses) du corpus ESLO 1, vu comme étape liminaire à ESLO 2, répondent à un objectif précis : participer à la réflexion sur l'évolution des modèles et des méthodes de constitution et d'exploitation des corpus oraux destinés à des finalités linguistiques.

1. Quelques considérations sur le linguiste et ses corpus

Sans oser une présentation épistémologique de la place des corpus en linguistique, nous souhaitons présenter quelques considérations sur l'usage des corpus en linguistique qui sont à l'origine du projet de constitution, d'exploitation et de diffusion des corpus ESLO.

1.1. Données attestées et situées *VS* masse de données

La linguistique connaît actuellement un bouleversement méthodologique amorcé il y a plus de 30 ans. Les possibilités offertes par le traitement automatique du langage et notamment les techniques d'exploitation des documents numériques ont permis des développements théoriques fondés sur l'exploitation de corpus, mettant ceux-ci aux centres de la description et de l'analyse linguistique.

Une ambiguïté demeure cependant. En effet, les corpus étaient utilisés bien avant le développement du domaine du TAL. Travailler sur corpus consistait alors à considérer l'objet d'étude comme une collection ordonnée de productions attestées et situées. Cette définition de l'objet impliquait une démarche empirique de description des faits qui s'opposait à une démarche hypothético-déductive fondée sur l'intuition du chercheur. La méthodologie de travail sur corpus

[1] Enquête Socio-Linguistique à Orléans.

était donc un acte scientifique fort et fondateur de certains domaines (sociolinguistique, analyse de la conversation, ethnolinguistique, etc.) centrés sur la conception de "corpus de langue parlée".
Depuis les années 1980, la linguistique de corpus s'est définie autour de grands corpus de langue écrite traités informatiquement comme l'ont décrit Kennedy (Kennedy, 1998) pour l'anglais et Habert pour le français (Habert *et al.*, 1997).
Ainsi les possibilités offertes par le traitement informatique de masse de données sont devenues l'atout principal de la linguistique de corpus. Toutefois trois questions, selon nous centrales, se trouvent biaisées dans ce contexte :

- le corpus est souvent constitué de productions très normées (romans, articles de presse, textes officiels) dont le traitement requiert une standardisation (orthographe, étiquetage, etc.) ;
- le corpus est souvent considéré comme représentatif d'une hétérogénéité des pratiques de par le simple fait qu'il constitue une masse de données ;
- la disponibilité de vastes corpus (FRANTEXT) permet dans de nombreux travaux d'éluder la question pourtant centrale de la constitution du corpus comme première étape d'une théorie linguistique.

1.2. Place des corpus oraux

Si la linguiste du corpus s'est massivement développée, force est de constater que la linguistique dispose de peu de corpus oraux. C'est un paramètre qu'il est facile d'expliquer : la tradition littérale est continue depuis l'Antiquité quand les modes de conservation du son ont moins d'un siècle et demi d'existence. Mais ce n'est pas l'unique raison. L'oral s'accommode beaucoup moins d'un traitement excluant les variations. L'écrit est normalisé par sa présentation même en chaîne de caractères. Il est le produit d'une transcription déjà effectuée, que la source en soit assignée au mental ou au signal. Avant tout retravail par les instruments et les outils du TAL, une homogénéisation de la présentation et des formes a été accomplie à divers niveaux : orthographe, découpe des mots et des phrases, ponctuation…
La recherche en intelligence artificielle a été facilitée, quand elle se donnait les langues pour objet, par la saisie d'énoncés écrits, avec pour conséquence l'élaboration de techniques et d'approches dont l'extension à des corpus oraux (enregistrements, transcriptions phonétiques) était malaisée. On est en présence d'un cas d'école concernant l'ajustement réciproque des données et des outils qui pâtit de l'extension des processus à de nouvelles catégories d'objets.
Les problèmes de l'extension des méthodes éprouvées sur des corpus scripturaux à des corpus oraux se situent sur différents plans :

- insuffisance des corpus oraux, que ce soit en termes quantitatifs de disponibilité globale, ou qualitatifs de fiabilité scientifique ou de prétraitement ;
- dissymétrie des champs d'application de l'enquête (opposition des études linguistiques de terrain - field linguistics), orientées vers les langues sans tradition écrite, et des linguistiques de bureau (armchair linguistics), centrées sur les textes de référence et l'écrit -, les départements informatiques étant plus souvent confrontés à celles-ci pour lesquelles existe de surcroît une forte demande des industries de la langue ;
- parcellisation des enquêtes et des standards retenus pour la collecte, la conservation et la codification : ainsi, le très important travail d'archive orale entrepris dans les deux dernières décennies par les historiens et les sociologues a souvent été entrepris sans finalité externe qui aurait pu assurer une exploitation ouverte des fonds ;
- faible exigence de prescription : les corpus sont constitués sur des objectifs *ad hoc*, ciblant leur finalité en fonction d'objectifs circonscrits, par exemple la reconnaissance vocale ou la fouille de textes ;
- pratiques lacunaires de catalogage et de description des ressources : la bibliothéconomie des archives sonores reste aujourd'hui encore balbutiante et c'est un chantier international où il importe que soient formulées des propositions pour tout ce qui a trait à la standardisation des produits, à l'indexation et à la consultation (représentativité des éléments de catalogage par rapport aux contenus en fonction de pertinences multiples).

1.3. Corpus disponibles *VS* corpus fantômes

Nous l'avons précisé, si la linguistique de corpus s'est considérablement développée, ce n'est pas pour autant, et le fait mérite d'être grassement souligné, que les corpus eux même soient disponibles. En effet, à l'exception notamment de FRANTEXT et du *Monde*, les corpus sont toujours évoqués dans les travaux, mais ne sont que très rarement diffusés. Ils jalonnent les articles et les thèses comme les fantômes hantent les couloirs et les tours : toujours évoqués comme preuve mais n'apparaissant à nul autre qu'à celui qui en parle.

Cette situation ne mérite pas d'être caricaturée et on peut esquisser une typologie des corpus en fonction d'un critère de disponibilité :

- Certains corpus ont été constitués dans le cadre d'une recherche précise et n'ont de pertinence que pour celle-ci. Les conditions de collecte ou le travail très spécifique d'annotation ne permet pas la diffusion de ces données.
- D'autres corpus ne sont pas disponibles par volonté des chercheurs qui souhaitent garder une priorité scientifique sur un travail de collecte coûteux et laborieux.
- Enfin il existe des corpus conçus comme des bases de données qui prennent le statut de corpus de référence par le simple fait que ce sont les seuls disponibles. Ces corpus sont alors utilisés simplement… parce qu'ils sont là.

Nous ne pouvons terminer cette courte typologie sans évoquer les corpus totalement fantômes qui fondent certains travaux sans qu'aucune information ne précise les raisons de l'absence de l'accès aux données, pourtant seule garantie d'un travail scientifique en principe ouvert à la falsification.

Le programme ESLO se situe résolument dans une démarche scientifique pour laquelle un corpus non disponible n'existe pas.

En bref, la linguistique de corpus a dans un premier temps peu pris en charge le domaine de la langue parlée et des données situées. Cependant les technologies récentes permettant de numériser le son et d'avoir une synchronisation temporelle entre le signal et une ou des transcriptions ainsi que les initiatives de normalisation de structuration des corpus (*TEI*), des métadonnées (*Dublin core* et *Olac* par exemple) et des données liées ouvrent de nouvelles perspectives pour la linguistique de corpus. Toutefois il n'existe pas actuellement de grand corpus français de langue parlée disponible pour la communauté scientifique. Le projet ESLO (cf. *infra*) souhaite répondre à cette demande.

2. Les corpus ESLO : de la collecte à l'exploitation d'un corpus oral

2.1. ESLO 1, un corpus à reconstruire

L'Enquête Socio-Linguistique à Orléans (désormais : ESLO 1) a été conduite en 1968 par des universitaires britanniques avec une visée didactique : l'enseignement du français langue étrangère dans le système public d'éducation anglais. Il s'agit d'un vaste corpus estimé à plus de 300 heures (environ 4 500 000 mots).

Elle comprend environ 200 interviews, toutes référencées (caractérisation sociologique des témoins, identification de l'enquêteur, date et lieu de passation de l'entretien), mais aussi une gamme d'enregistrements variés (conversations téléphoniques, réunions publiques, transactions commerciales, repas de famille, entretiens médico-pédagogiques, etc.). Certains des enquêtés ont ainsi été enregistrés dans des situations très différentes. ESLO 1 couvre l'ensemble des catégories socioprofessionnelles, hommes et femmes, avec plusieurs locuteurs originaires de différentes régions. C'est un échantillon des formats de la communication, des tâches linguistiques, des types de discours selon une approche essentiellement dialogique. Ce corpus représente, par son ampleur, sa rigueur et sa cohérence, le plus important témoignage disponible sur le français parlé avant 1980. Si les fins de sa constitution étaient linguistiques, ESLO 1 est un témoignage unique sur les jugements concernant mai 68 vu de la province ou sur les représentations collectives de la cité à cette époque.

Le Coral (Centre orléanais de recherche en anthropologie et linguistique) a réussi à récupérer en 1993 l'ensemble de documents originaux composés des bandes magnétiques, un catalogue dactylographié, quelques centaines de feuillets de transcription manuscrites (d'une qualité inégale) et les fiches d'identification des locuteurs.

L'opportunité offerte par la numérisation des originaux arrivés en fin de vie a permis au Coral de consacrer un projet à la conservation et à la valorisation du corpus. L'opération de numérisation n'était pas en l'occurrence anodine ; c'est une véritable reconstruction du corpus et sa transformation en un nouvel objet scientifique qui a été opérée. Les documents sonores ont été

récolligés et complétés (la conservation avait été défectueuse), numérisés à partir des enregistrements et une indexation et un premier catalogage informatisé a pu être réalisé. Parallèlement, l'exploitation exhaustive d'un sous-ensemble a été entreprise au point de rencontre de données linguistiques variationnistes et cognitives (description d'une tâche). L'étape suivante consiste à transcrire et baliser l'intégralité du corpus.
L'enjeu de cette reconstruction n'est pas neutre. Il s'agit d'établir des principes ayant valeur de normalisation afin de mettre l'ensemble des données à la disposition de la communauté scientifique dans un format qui en permette une exploitation fiable, optimale et intensive, y compris pour des applications industrielles après sélection des contenus.

2.2. ESLO 2, un corpus à anticiper

En partant des acquis d'ESLO 1, une nouvelle enquête, dénommée ESLO 2, a été mise en chantier par le CORAL. Il s'agit, à quarante années de distance, de constituer un corpus comparable dans le produit attendu et dans les modalités de la collecte : l'objectif a été fixé à 400 heures environ de documents sonores qui totaliseraient approximativement 6 000 000 de mots. Réunis, ESLO 1 et ESLO 2 formeront une collection de 700 heures d'enregistrement, soit plus de 10 000 000 de mots, ce qui est considéré aujourd'hui comme une valeur repère pour les investigations projetées.
ESLO 2 a été conçu pour préfigurer la référence attendue dans un domaine qui en est encore à se structurer et dans lequel se manifeste de manière récurrente une demande de définition pour un format standardisé de *collecte*, de *conservation*, de *traitement* et d'*analyse* :
- la *collecte* sur le terrain est première, non seulement dans ses aspects techniques, aujourd'hui bien maîtrisés, mais dans la définition du profil de l'échantillon représentatif et dans la problématisation des interactions entre les témoins et les enquêteurs ;
- la *conservation*, qui inclut la préservation des supports, l'indexation des contenus et l'accessibilité (c'est-à-dire la protection) des données, conditionne le partage des sources à des fins d'étude scientifique ou didactique ;
- le *traitement*, en lien étroit avec le développement des matériels et des langages informatiques, suppose la maîtrise d'une chaîne d'opérations, depuis la conversion numérique des enregistrements jusqu'à une transcription balisée et ouverte à l'ensemble des interrogations pertinentes pour les demandes du linguiste, du sociologue ou des décideurs, des didacticiens voire du grand public ;
- l'*analyse* constitue l'épreuve des théories (et des logiciels) puisqu'elle compare les formalisations et les opérations et qu'elle valide ou infirme les hypothèses en prenant argument de leur compatibilité aux faits.
Les acquis en matière de conservation, de traitement et d'analyse seront reportés sur ESLO 1 comme le requiert la comparabilité attendue.

3. Choix pour la mutualisation et l'interopérabilité d'un grand corpus oral

Quels sont les choix et les enjeux contenus dans l'objectif de mutualisation et d'interopérabilité d'un grand corpus oral de type ESLO ? Nous nous bornerons ici à présenter une démarche suffisamment générale qui interroge l'exploitation des corpus en sciences humaines.

3.1. Rôle des métadonnées pour l'interopérabilité

Les corpus constituent des ressources numériques sur lesquelles se fondent la majorité des travaux en linguistique actuellement sans que la question de la forme de ceux-ci et des limites qui bornent cette collection ordonnée soit toujours clairement résolue.
Un corpus est constitué de données brutes et/ou annotées (Véronis 2000, Habert et Fuchs 2004). Dans le cas des corpus oraux, les enregistrements de la parole constituent les données primaires, la transcription et les autres annotations éventuelles représentent des données secondaires. L'ensemble de ces données sont décrites par des métadonnées chargées de documenter le corpus.
Ce sont ces dernières informations, particulièrement importantes pour rendre une ressource disponible mais aussi pour expliciter les critères de sélection et d'organisation des données et donc des bornes du corpus, qui manquent souvent dans les corpus disponibles. Or il s'agit ni plus ni moins de poser ainsi la question de la représentativité du corpus. C'est en effet l'explicitation des bornes du corpus (conditions de production, de réception, contexte des usages, informations

sociologiques sur les producteurs, genre, etc.) qui permet de juger de la représentativité du corpus qui du statut d'échantillon de la langue passe très souvent à celui de corpus de référence (même si ce statut référentiel est implicite) sans aucun regard réflexif sur la forme de celui-ci.
Dans le cas du corpus ESLO 1, l'équipe a souhaité conserver le travail de catalogage et de documentation déjà anticipé par les auteurs du corpus. La démarche actuelle et validée par la communauté consiste à utiliser des métadonnées *Dublin Core* et les extensions préconisées par le programme OLAC. Ce jeu d'étiquettes permet des opérations de catalogage tout à fait satisfaisantes en termes de description d'une ressource qu'on souhaite répertorier pour la rendre accessible.
Cependant cette procédure n'est pas suffisante pour documenter un corpus conçu comme un réservoir qui doit permettre à un chercheur de construire son propre corpus répondant aux exigences de sa recherche. Dans le cadre de cette extraction/construction, les informations permettant de borner le corpus doivent répondre à une granularité très fine. Dans le cas du corpus des ESLO, nous avons déjà pointé l'importance accordée aux informations sur les locuteurs, la situation de collecte et l'échantillonnage dont le but était de constituer un corpus représentatif.
Ce travail méthodologique permet de considérer que la collecte a correspondu à un travail méthodologique rigoureux, source de données représentatives et qu'ainsi le chercheur est sûr d'être confronté à des données pertinentes. Il convient d'aller plus loin et de considérer qu'un corpus doit contenir une riche documentation sur les données mais aussi sur les contextes de production de ces données. Ces contextes concernent aussi bien les données sur les locuteurs et la situation de collecte que l'explicitation de la démarche du chercheur.
Quelles sont les possibilités pour mettre à disposition un corpus qui contient en lui-même les informations sur ses bornes constitutives ? Le standard XML offre des éléments de réponse.
Techniquement ce standard sépare la représentation physique et logique des documents (les données et les métadonnées). Tout document XML comporte donc l'identification des éléments possibles et leurs relations possibles (Définition de Type de Document) et les données identifiées selon cette DTD. C'est alors la notion même de données brutes qui est redéfinie. Ainsi la TEI rend obligatoire la constitution d'un header (en-tête) en début de corpus qui recense les informations sur le contexte de production des données. Cependant le chapitre de la TEI consacré à l'oral est actuellement beaucoup trop succinct pour permettre une véritable normalisation de cette démarche.
Il y a donc un enjeu à considérer les métadonnées comme des éléments de description des données au sens linguistique de celle-ci et non simplement en termes de documentation de ressources. Les métadonnées doivent permettre d'expliciter la démarche du chercheur en proposant une description fine de ses choix théoriques "encapsulés" dans des choix techniques. Les opérations de transcription sont en ce sens un exemple particulièrement éclairant.

3.2. Des données exploitables : le cas de la transcription

La difficulté la plus importante rencontrée par les initiateurs d'ESLO 1 a été l'ampleur de la tâche de transcription. Sur ce point aussi, et même principalement, l'avancée technologique bouleverse l'objet scientifique.
Depuis quelques années, alors que la manipulation du son numérique devenait très aisée (capacité de stockage, rapidité d'accès, débit suffisant pour une transmission en réseau…), des logiciels permettent la synchronisation du son et de la transcription (*Praat*, *Transcriber*, *Winpitch*, *soundedit*, etc.).
Ces innovations ont des répercussions méthodologiques importantes sur le travail du linguiste. En effet, avec des transcriptions alignées sur le signal sonore, l'oral devient physiquement l'objet d'étude et est systématiquement disponible en même temps que la transcription. Le retour aux données peut alors être systématique, ce qui est de nature à faciliter les procédures de vérification, étape essentielle du travail scientifique, malheureusement souvent rendue impraticable de par l'inaccessibilité des corpus.
Parallèlement, la synchronisation, qui permet l'annotation de segments temporels, offre une base de référence pour de la multi annotation et donc de la multi transcription. On peut concevoir, pour un même segment, une multitude de transcriptions, opérées dans des cadres théoriques distincts et/ou avec des granularités différentes, dont chacune répond à un besoin scientifique spécifique. Ici, la transcription n'est plus la vérité d'un chercheur (au mieux) ou d'un transcripteur, elle devient cumulative.

Face à l'ampleur de la tâche, les choix pour la transcription d'ESLO 1 ont été fondés sur la volonté de mettre à disposition une transcription de l'intégralité du corpus le plus rapidement possible sans que celle-ci n'implique une théorie linguistique très déterminée (même si toute transcription est une formalisation impliquant une théorie).
Cette première transcription est conçue comme une transcription de base avec un simple statut d'outil de navigation au sein du corpus sonore et de repérage de phénomènes selon une granularité grossière. L'outil sélectionné a été *Transcriber* pour sa simplicité d'utilisation, sa robustesse face à des fichiers longs, et sa sortie en un format de fichier XML qui nous a semblé être une garantie d'interopérabilité.
Les conventions de transcription ont donc été réduites au minimum : la segmentation se fait sur une unité intuitive de type "groupe de souffle et/ou unité syntaxique pertinente". Le tour de parole a été défini par les changements de locuteurs uniquement, les pauses indiquées automatiquement par leur durée (précision du centième de seconde).

3.3. Corpus mutualisé pour des analyses multi-domaines : le test d'eslomelette

Le groupe du CORAL qui travaille sur ESLO est composé de chercheurs dont les sensibilités théoriques sont diverses, et les domaines de compétence en linguistique assez variés, allant de l'épistémologie à la syntaxe, en passant par le TAL, la phonologie, la pragmatique, la sociolinguistique, etc.
Cette diversité théorique est vue comme un atout majeur pour ce projet, dont l'objectif premier est la mise à disposition d'un corpus, laquelle ne peut pas être conçue sans multi-opérabilité.
Or, quelle meilleure garantie pour un corpus qui se veut disponible pour la communauté linguistique dans son ensemble que d'être conçu par des chercheurs dont les centres d'intérêts sont assez divers ?
Pour mettre à l'épreuve cette première piste, l'équipe a choisi un échantillon de ce corpus sur lequel elle a décidé d'opérer toutes les étapes du travail linguistique, de l'identification du corpus jusqu'à l'analyse, opérée dans des domaines divers (syntaxe, pragmatique, lexique, phonologie…) , en passant par l'annotation, qui comporte elle-même différentes phases (transcription, annotation, métadonnées). L'équipe a donc constitué un sous-corpus composé des 90 réponses à la question "comment-faites-vous une omelette ?". Les couples questions réponses ont été transcrits selon les conventions testées. Ces fichiers de transcription ainsi que l'ensemble des métadonnées constituent une collection de documents intégrés à une base de donnée XML native. Une interface (xquery) a été réalisée dans le cadre du projet GRICO[1] et du CRDO[2] après un travail conjoint d'informaticiens spécialisés dans la gestion de corpus oraux et les chercheurs en linguistique de l'équipe.
Cette première expérience est intéressante à plus d'un égard. D'abord, elle permet de voir sur un petit échantillon toutes les erreurs (d'annotation, de structuration), qu'il est encore temps d'éviter pour la totalité du corpus. Ensuite, elle précise l'utilité d'un corpus situé.
Pour ne donner ici qu'un exemple, on peut citer une recherche en pragmatique opérée par des membres de l'équipe. L'analyse pragmatique de la question de l'omelette montre qu'à partir de la question zéro, telle qu'elle figure dans le questionnaire – *i.e.* « Comment est-ce qu'on fait une omelette ? Pourriez-vous m'expliquer comment on fait ? » – les enquêteurs, visiblement gênés par la question, développent toutes sortes de modalisation. Après le relevé systématique des différentes marques de distanciation vis-à-vis de la question, qui se distinguent en fait en deux groupes, à savoir d'une part les « stratégies de justification » (évocation des écarts culturels entre la France et l'Angleterre, contrôle de la qualité du son, etc.) et, d'autre part, les « stratégies d'atténuation » (emploi du conditionnel, l'enchâssement de la question, l'emploi de l'atténuation autonymique, etc.), on peut se poser une série de questions que la nature et la structuration du corpus permettent, et qui auraient été tout simplement impossibles ailleurs. Par exemple, y a-t-il dans ce dégradé de modalisation une variable sociologique ? Autrement dit, l'enquêteur utilise-t-il plus ou moins de modalisation selon le profil de l'enquêté (son âge, son sexe, son niveau sur l'échelle AM) ? Ce type de questions, combien intéressantes d'un point de vue linguistique, est tout simplement impossible dans d'autres corpus. Autre interrogation : y a-t-il une variable individuelle ? Autrement dit, les enquêteurs se distinguent-ils les uns des autres vis-à-vis de leur relation avec la

[1] Groupe de Recherche sur l'Interopérabilité des Corpus Oraux. Michel Jacobson (Lacito-CRDO) et Richard Walter (Modyco).
[2] Centre de Ressources pour la Description de l'Oral. http://crdo.vjf.cnrs.fr:8080/exist/crdo/

question ? Et, d'ailleurs, un enquêteur quelconque utilise-t-il au fil du temps que dure l'enquête (en l'occurrence presque un an) les mêmes stratégies de modalisation ? Toutes ces interrogations, et bien d'autres, auraient été fastidieuses ailleurs : ici, elles sont non seulement possibles, grâce à la fois aux outils du TAL et à la disponibilité des métadonnées, mais en plus utiles : par exemple, les interrogations naïves qui viennent d'être évoquées permettent de poser des questions cruciales, concernant la réflexivité de l'enquête, son statut, son degré de figement et d'interaction, etc. Derrière ces questions, il s'agit ni plus ni moins de poser la question de la pertinence et de la validité de données non situées.
Cet enjeu n'est pas restreint à la pragmatique et à la sociolinguistique, le travail entrepris par des chercheurs aux objectifs très différents permet de tester les possibilités de réappropriation de contraintes méthodologiques (par exemple, la normalisation recherchée par le chercheur en TAL est-elle compatible avec le linguiste variationniste ?).

4. Conclusion

Les enjeux inhérents à l'exploitation d'un grand corpus oral ne se résument pas à des choix techniques imposés par les outils du traitement automatique du langage et de la linguistique de corpus. L'exemple des corpus d'ESLO ne met en évidence que ce qu'on savait déjà : "on ne peut dissocier l'accumulation des données et la critique de leur constitution".
Cette évidence interroge la linguistique sur la constitution même de son objet mais aussi l'ensemble des sciences sociales sur l'exploitation de la masse de données. La réponse passe nécessairement par la maîtrise de la totalité de la chaîne : de la collecte des données à leur organisation à des fins d'analyses variées.

BIBLIOGRAPHIE

ABOUDA, L. 2004. Deux types d'imparfait atténuatif, *Langue française*, 142, pp. 58-74.

BAUDE, O., JACOBSON, M., TCHOBANOV, A., et WALTER, R. (à paraître) Interopérabilité des corpus sonores : le cas des corpus en français, *Colloque international Phonological variation : the case of French*, 25-27 août 2005, Tromso.

BAUDE, O. 2004. Les corpus oraux entre science et patrimoine. L'expérience de l'observatoire des pratiques linguistiques, in *Actes du Colloque international du GRESEC « La publicisation de la science »*, Grenoble, pp. 7-11.

BAUDE, O. (éd.) 2006. *Corpus oraux. Guide des bonnes pratiques 2006*, Paris, Cnrs éditions – Orléans, PUO.

BERGOUNIOUX, G. 1992, « Les enquêtes de terrain en France », *Langue française*, 93, pp. 3-21.

BERGOUNIOUX, G., BARADUC, J., DUMONT, C. 1992. L'Etude socio-linguistique sur Orléans (1966-1991), 25 ans d'histoire d'un corpus, *Langue française*, 93, pp. 74-93.

BLANCHE-BENVENISTE, C., JEANJEAN, C. 1987. *Le français parlé, transcription et édition*, Paris, Didier érudition.

BLANC, M., BIGGS, P. 1971. L'enquête sociolinguistique sur le français parlé à Orléans, *Le français dans le monde*, 85, pp. 16-25.

DELAIS-ROUSSARIE, E. et DURAND, J. (éds.) 2003. *Corpus et variation en phonologie du français, méthodes et analyses*, Toulouse, PUM.

EAGLES, 1996, Preliminary Recommendations on Spoken Texts, EAG-TCWG-SPT/P, Pise, Consiglio Nazionale delle Ricerche, Instituto di Linguistica Computazionale.

HABERT, B. *et al.* 1997. *Les linguistiques de corpus*, Paris, Armand Colin.

HABERT, B. et Fuchs, C. 2004. Introduction le traitement automatique des langues : des modèles aux ressources, *Le français moderne traitement automatique et ressources numérisées pour le français*, pp. 1-13.

MERTENS, P. 2002. Les corpus de français parlé ELICOP : consultation et exploitation, in J. Binon *et al.* (éds.), *Tableaux Vivants*, Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock, Leuven, Universitare Pers.

PIERREL, J.-M. (éd.) 2000. *Ingénierie des langues*, Paris, Hermès sciences.

RASTIER, F. 2004. Enjeux épistémologiques de la linguistique de corpus, *Texto !* [en ligne], juin 2004. Rubrique Dits et inédits. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html>.

SINCLAIR, J. 1996. Preliminary recommendations on corpus Typology, Technical Report, Eagles.

« Speech Annotation And Corpus Tools », A special issue of Speech Communication Volume 33, numbers 1-2, 2001, Edited by Steven Bird and Jonathan Harrington.
VERONIS, J. 2000. Annotation automatique de corpus : panorama et état de la technique, in J.-M. Pierrel (éd.), pp 111-130.
WYNNE, M. 2005. *Developing Linguistic Corpora : a Guide to Good Practice*, AHDS, http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm. Visité le 01 juillet 2006.

LA LINGUISTIQUE ET LE CORPUS : UNE AFFAIRE PRÉPOSITIONNELLE

Geoffrey WILLIAMS
Université de Bretagne Sud, Lorient

SOMMAIRE

Résumé : *La linguistique de corpus a été très largement développée comme discipline dans le monde anglo-saxon. Ce paradigme de recherche est sorti de la linguistique appliquée à partir de deux grandes traditions ; l'enseignement de l'anglais comme langue seconde et une approche contextualiste de la linguistique, approche associée avec Firth. Dans cette communication, je montre comment les deux traditions se sont fusionnées avec le projet COBUILD. Je décris l'évolution de la discipline comme paradigme de recherche développé autour des corpus soigneusement constitués et utilisant une analyse inductive. Dans la conclusion, je plaide pour la reconnaissance de la linguistique de corpus autonome par opposition à la linguistique sur corpus qui implique d'autres disciplines telles que la sociolinguistique ou le TAL.*

Abstract : Corpus Linguistics is widely developed in the English-speaking world. The research paradigm has developed within applied linguistics out of two related strands in English language studies; the teaching of English as a second language and the contextualist approach to linguistics associated with Firth. This paper seeks to show how the two traditions developed before merging through the work done on the COBUILD project. It then illustrates how contextualist corpus linguistics has developed as a research paradigm in which a carefully constructed corpus built following linguistic criteria is explored using a corpus-driven approach. The paper concludes by claiming that although disciplines in linguistics and natural language processing may work on corpora, they may not be doing corpus linguistics, which remains an independent discipline with its own methodology and theoretical stance.

1. Introduction

En anglais, la situation est simple, '*corpus linguistics'* est un mot composé formé de deux substantifs, dont l'un va limiter le champ de référence de l'autre. La linguistique est une discipline, le mot corpus décrit l'objet. Le reste est sujet à interprétation, la puissance de l'anglais est dans l'ambiguïté, une ambiguïté que nous n'essaierons pas de lever dans l'immédiat.

Pour le français, la situation est plus complexe puisque nous ne pouvons pas simplement juxtaposer deux mots, il faut les lier avec une préposition, et le choix d'une préposition implique une interprétation. Faut-il mettre *de* ou *des*, ou peut-être *sur* ?

Avant même de choisir la préposition, nous rencontrons une difficulté supplémentaire : le mot '*linguistics'* en anglais semble être un pluriel, mais comme '*physics'* ou '*mathematics'*, il est en réalité invariable. Par conséquent, 'de' signifiera la présence d'une discipline unique, 'des', que plusieurs disciplines - au lieu de plusieurs approches de la même discipline - sont en jeu. 'Sur' est une interprétation supplémentaire impliquant que d'autre domaines de la linguistique peuvent utiliser les corpus sans faire de la linguistique de corpus *per se*, ce qui soulève la question de la

nature des corpus.
Le but de cette communication est d'essayer de démêler les différentes interprétations de '*corpus linguistics*' en décrivant l'origine anglo-saxonne de la discipline, le contexte de recherche de la discipline de son origine à nos jours. En comparant les différentes interprétations françaises du terme nous essaierons non d'imposer une définition, mais de clarifier la situation entre les différentes approches de cette discipline.

2. Les origines de la linguistique de corpus dans la tradition anglo-saxonne

Il est toujours trop facile d'essayer de trouver un inventeur, comme s'il suffisait de crier Euréka et de trouver des merveilles. Cette tendance est exaspérée par des velléités patriotiques : la théorie de Darwin a provoqué un intérêt pour les écrits de Lamarck, décriés auparavant, la parenté de la photographie est disputée entre un Daguerre et un Fox Talbot, et ainsi de suite.
En ce qui concerne la linguistique de corpus, la question de l'antériorité des corpus se dispute entre plusieurs corpus électroniques, FRANTEXT, Brown, OSTI… En réalité, chaque ensemble textuel a été créé en reconnaissant des possibilités offertes par l'informatique naissante afin de résoudre des problématiques différentes. Il est donc inutile de chercher l'antériorité d'un tel ou un tel, d'autant plus que, pendant cette période antérieure au courrier électronique, les chercheurs travaillaient en relative isolation. Le plus important sera de voir pourquoi et comment des tendances actuelles ont évolué afin que des chercheurs de nos jours puissent échanger des informations en comprenant l'autre.
Dans la tradition anglo-saxonne de la linguistique de corpus, la lexicographie, l'enseignement et les corpus sont intimement liés. La tendance contextualiste est le fruit de l'interaction entre les trois éléments de base.

2.1. La situation avant 1945

2.1.1. La lexicographie

On peut ainsi dater le développement de la linguistique de corpus à 1755 avec le dictionnaire de Johnson, le premier dictionnaire basé sur un 'corpus' sous la forme de fiches de travail accompagnées de citations. Une telle affirmation est peut-être un peu osée, mais pas totalement infondée puisqu'avec Johnson débute une tradition lexicographique plus normative que prescriptive mais basée sur des textes authentiques, bien que limitée à des textes 'nobles' de la littérature. La tradition lexicographique instaurée par Johnson est à la base du *Oxford English Dictionary*. Plus récemment, la tradition lexicographique d'Oxford a donné naissance à une autre forme de dictionnaire, le dictionnaire pour apprenant, avec le *Oxford Advanced Learner's Dictionary*, issu du *Learner's Dictionary of Current English* de Hornby publié en 1948 (Cowie). Ces dictionnaires pour apprenants sont toujours basés sur des fiches, mais avec des exemples tirés de la langue générale. Le ton a changé en 1987 avec la publication du *COBUILD Advanced Learner's English Dictionary*, basé sur un grand corpus de référence. Dorénavant tous les dictionnaires pour apprenants seront basés sur corpus, et les corpus seront de plus en plus utilisés pour l'élaboration de dictionnaires monolingue et bilingue des éditeurs britanniques, et maintenant dans beaucoup d'autres pays. Ce qui a poussé à ces deux révolutions, celle de Hornby, puis celle de l'équipe de COBUILD, est l'enseignement de l'anglais comme langue seconde.

2.1.2. L'enseignement de l'anglais comme langue seconde

Grâce à l'Empire Britannique, l'anglais était devenu dans la période suivant la première guerre mondiale une langue dominante dans les affaires. Il fallait par conséquent que les gens apprennent l'anglais (pas nécessairement celui de la langue de Shakespeare) d'une manière plus pragmatique pour le travail. Les bases pour un enseignement de la langue fondé sur une linguistique appliquée avaient déjà été jetées avec la publication de Sweet's 'Practical study of Langages' en 1899 (Howatt 1984), développé à partir d'un article publié en 1884. Dans l'approche de Sweet, le lexique et la phraséologie étaient centraux, mais il fallait que le lexique soit structuré et que les phrases soient un lien entre le texte et la grammaire, autrement dit, un certain contexte était nécessaire pour apprendre. Les phrases ne seront pas inventées, mais authentiques, l'autre credo du contextualisme.
L'enseignement des langues s'est beaucoup développé dans la période avant la première guerre mondiale, mais en ce qui concerne la linguistique de corpus, la période la plus importante date de l'entre-deux-guerres avec les travaux de Palmer au Japon. Cette période a vu un intérêt intense

pour des vocabulaires essentiels pour apprenants, mais également les premiers travaux sur la collocation en anglais.
Pendant ses années au Japon, Palmer a publié extensivement sur la théorie et la pratique de l'enseignement de l'anglais comme langue seconde (Howatt op.cit). Palmer s'est beaucoup investi dans l'étude du lexique, dans le but de créer un vocabulaire contrôlé pour l'apprentissage, deux rapports ayant été publiés sur ce thème. Il a aussi collaboré avec West, l'auteur du '*General Service List*', liste de mots à la base de nombreuses méthodes d'apprentissage. C'est précisément cet intérêt pour un vocabulaire restreint au service des apprenants qui a donné naissance à un dictionnaire de langue générale pour apprenants, le *Learner's Dictionary of Current English* de Hornby.
L'autre aspect des travaux sur le vocabulaire de Palmer est son rapport sur les collocations, *Second Interim Report on English Collocations* (Palmer 1933). L'étude des collocations était une suite logique à des rapports sur le vocabulaire montrant qu'au delà des mots simples, il y avait ce que Palmer a appelé des « *comings-together-of-words* », des rassemblements de mots (*ibid*. p.1). Après une discussion des classifications possibles, Palmer décide de les appeler 'collocations', réutilisant un terme vague ayant déjà été employé par Sweet. D'après Palmer, il sera nécessaire de définir ce que l'apprenant doit apprendre comme combinaisons ; les combinaisons figées et sémi-figées. La suite est une classification des collocations par partie de discours. Ces collocations sont trouvées dans des textes authentiques, mais par le biais de l'intuition du linguiste.
Le rapport est souvent cité, mais n'a jamais été largement publié. La tradition collocationnelle de Palmer a beaucoup influencé la phraséologie, tradition qui a cependant largement ignoré les possibilités offertes par les corpus jusqu'assez récemment. L'analyse des collocations en corpus est issue d'une autre tradition de recherche, le contextualisme de Firth.

2.1.3. Firth et le contextualism

Firth est souvent vu comme le père de la collocation, même si ses écrits sont postérieurs à ceux de Palmer. Il est probable que nous ayons ici une des coïncidences historiques de découvertes quasi-simultanées. Il est possible que les travaux de Firth soient aussi plus largement lus en raison de sa position de Professeur de linguistique à Londres et de la large diffusion de ses écrits par ses étudiants. Les écrits de Firth sont beaucoup plus énigmatiques que ceux de Palmer, sans la démonstration pratique que nous donne le 'Interim Report'. La phrase célèbre de Firth « you shall know a word from the company it keeps » montre que le point de vue est différent de celui de Palmer. Pour celui-ci, il s'agissait d'unités polylexicales à découvrir, à mettre dans un dictionnaire et à transmettre aux apprenants, mais « the company words keep » est une approche autre, où la nécessité d'avoir des ensembles bien formés est moins importante que la notion d'associativité. La différence se trouve dans une approche textuelle, par opposition à une approche lexicographique, de la collocation. La textualité est centrale aux thèses de Firth qui ont développé les notions de contexte de culture et contexte de situation de Malinowski.
Anthropologue de renom, Malinowski reste très connu pour ses travaux sur les habitants des îles Trobriand. Il a reconnu très tôt l'importance de prendre en compte les aspects culturels dans la compréhension de la langue, le sens ne pouvant pas être évalué en dehors du contexte de situation.

> *Without some imperative stimulus of the moment, there can be no spoken statement. In each case, therefore, utterance and situation are bound up inextricably with each other and the context of situation is indispensable for the understanding of the words* (Malinoswki 1924. 307).

Ces deux notions de base ont été reprises et développées par Firth, qui a travaillé également à l'Université de Londres, pour élaborer une théorie linguistique ; le contextualisme. La linguistique de Firth était un rejet de l'approche mentaliste. Selon lui (1935 : 19)

> *I do not therefore follow Ogden and Richards in regarding meaning as relations in a hidden mental process, but chiefly as situational relations in a context of situation and in that kind of language which disturbs the air and other people's ears, as modes of behaviour in relation to the other elements in the context of situation*

Firth était néanmoins un homme de son époque, ses sources sont authentiques, mais largement littéraires. Firth est resté aussi un théoricien du langage, le contextualisme ayant surtout été développé par ses étudiants, notamment Halliday et Sinclair.

2.2. Le contextualisme d'après guerre

Dans le développement du contextualisme, deux disciples de Firth sont à noter : Halliday et

Sinclair. Halliday est à l'origine de la grammaire systémique et fonctionnelle, une grammaire descriptive très employée dans la linguistique de corpus contextualiste puisque complète, mais neutre. Si Halliday a surtout développé l'aspect grammatical, c'est Sinclair qui sera à l'origine de la partie lexicale et donc 'l'inventeur' de l'analyse de corpus contextualiste.
Une publication majeure dans le développement du contextualisme est parue en 1966, « In Memory of J. R. Firth » (Bazell et al.). Cette collection d'articles est à la fois une rétrospective sur les travaux de Firth, mort en 1960, et un programme pour le futur. Ainsi, des linguistes comme Jakobson et Lyons vont commenter l'apport de Firth, tandis que les articles de Halliday « Lexis as a linguistic level » et Sinclair « Beginning the study of lexis » annoncent les recherches qui vont mener à la grammaire systémique et fonctionnelle et à la linguistique de corpus contextualiste.

2.2.1. Halliday et la grammaire systémique et fonctionnelle

La théorie de Halliday a été annoncée dans son article de 1961 sur la catégorisation dans la grammaire. C'est une grammaire descriptive, textuelle et fermement basée sur le contexte. Ainsi, dans l'introduction de son œuvre majeure « An Introduction to Functional Grammar » (1994), il déclare que

> Just as each text has its environment, the 'context of situation' in Malinowski's terms, so the overall language system has its environment, Malinowski's 'context of culture'. The context of culture determines the nature of the code. As a language is manifested through its texts, a culture is manifested through its situations; so by attending to text-in-situation a child construes the code, and by using the code to interpret text he construes the culture. (1985 : xxxi)

Dans sa grammaire l'analyse est essentiellement descendante, du texte à la phrase, de la phrase aux mots. Cependant, dans une théorie de lexico grammaire, il y a forcement interaction entre la grammaire et le lexis. Ainsi il insiste que :

> A text is a semantic unit, not a grammatical one. But meanings are realized through wordings; and without a theory of wordings -- that is, a grammar -- there is no way of making one's interpretation of the meaning of a text. (ibid. xvii)

Dans son texte de 1966 annonçant le programme de recherche lexicale dans la grammaire, Halliday insiste sur le fait que la lexis est partie intégrante de la grammaire et constitue la partie la plus délicate, dans le sens de la plus fine, 'one-member classes' (1966 :149). Le fait que la lexis entre dans une classe unique ne veut pas dire que les mots sont relégués à une simple liste en marge de la grammaire. La grammaire de Halliday est systémique et multi-niveaux, il y a forcément une interaction entre tous les constituants qui forment le texte, et entre le texte et son environnement. Ainsi, la cohésion textuelle tient un rôle essentiel dans la grammaire (Halliday & Hasan 1971). Une partie de la notion de cohésion est basée sur la collocation, l'interaction entre mots. Tandis que Halliday utilise l'interaction collocationnelle dans le texte, Hoey l'a amenée plus loin dans le corpus (Hoey 1991, 2005).
En tant que grammaire descriptive, la grammaire systémique et fonctionnelle occupe une place de choix dans l'étude des corpus. Cependant, c'est largement une grammaire textuelle, l'aspect lexical ayant été traité par l'autre disciple de Firth, John Sinclair.

2.2.2. Sinclair et le rapport OSTI

Dans le titre même de son article en mémoire de Firth (1966), Sinclair a noté que nous n'étions qu'au début d'une étude contextualiste du lexique. Il a rapidement trouvé que l'outil informatique pouvait offrir un moyen d'aller plus loin. Ainsi il était amené à créer un corpus électronique. Le résultat de ces études sur corpus était un rapport publié en 1970, rapport qui a jeté les bases de la linguistique de corpus contextualiste, bien que peu diffusé à l'époque et publié seulement très récemment (Sinclair et al, 1970, 2004).
Le débat sur qui a créé le premier corpus électronique est largement stérile. Le mouvement vers une analyse des textes avec des outils informatiques était inévitable : il était dans l'air du temps, mais avec des objectifs différents. Comme l'a montré Léon (2005), l'arrivée de la théorie générative n'a eu aucun effet sur le développement de la linguistique de corpus contextualiste, qui a continué à évoluer dans le contexte de la linguistique appliquée.
Les premiers corpus ont été construits pour des raisons très différentes ; le TLF était largement littéraire, le Brown était également un corpus d'écrit, mais basé sur des échantillons et le *Survey of English Usage*, créé pour des recherches sur la syntaxe était largement inspiré par la tradition Firthienne mais n'a été numérisé que très tardivement. L'objectif du corpus OSTI était par contre d'explorer la lexis dans le paradigme contextualiste en faisant un corpus initialement basé sur

l'oral. Le projet a démarré en 1963 (Teubert, 2004). L'assemblage du corpus a commencé à l'Université d'Edimbourg et a été complété à l'Université de Birmingham. À l'époque, le fait d'avoir un ordinateur dédié à un projet linguistique était quelque chose d'extraordinaire dans un monde où uniquement les élites des sciences dures y avaient accès (Sinclair, communication personnelle).
Le rapport OSTI, officiellement '*The Report to the Office for Scientific and Technical Information (OSTI) on the Lexis Research project for the period January 1967 – September 1969'* était le résultat des travaux sur le corpus construit à Edimbourg et exploité à Birmingham. Outre la problématique de la création d'un corpus, le rapport est un véritable programme de recherche contextualiste, où les collocations s'avèrent centrales à l'approche. La notion de collocation significative a déjà été introduite par Sinclair (1966), mais ici la notion est explorée en relation avec des données issues du corpus. C'est dans ce rapport que les termes clés, comme *empan* et *fenêtre*, sont introduits et justifiés. Déjà la notion du principe d'idiome commence à apparaître.
Bizarrement, le rapport OSTI a été oublié par la suite, de la même manière que Palmer (1933) est souvent cité, mais n'est pas disponible. Néanmoins, l'approche élaborée dans le rapport OSTI a servi de base pour un projet encore plus ambitieux, le projet COBUILD.

2.2.3. L'école de Birmingham à COBUILD

COBUILD était une collaboration entre l'Université de Birmingham et les dictionnaires Collins. L'objectif était de construire un grand corpus de référence pour l'anglais et de l'utiliser pour la création d'un dictionnaire pour apprenants basé uniquement sur une analyse de corpus. C'est effectivement avec le projet COBUILD que nous trouvons unifiées les deux traditions d'étude de la collocation : la tradition de Palmer a été fructifiée dans *l'Oxford Advanced Learner's Dictionary,* et la tradition contextualiste s'est développée séparément. Avec le COBUILD, nous avons enfin un dictionnaire où la collocation trouve sa juste place, mais au lieu d'être basés sur l'intuition d'un lexicographe, les collocations et les sens doivent être justifiés par les données du corpus. Dans l'école de Birmingham, le rêve de Firth de voir la linguistique et la lexicographie unifiées a également été réalisé.
Le projet COBUILD était plus qu'un dictionnaire et un corpus. La création et l'exploitation du corpus ont été décrites par les membres de l'équipe (Sinclair et al. 1987). Mais de nombreuses autres applications sont issues de ce projet : des grammaires, des méthodes d'apprentissage, des études linguistiques... Les autres éditeurs de dictionnaires d'apprentissage ont été obligés de suivre, c'est ainsi que le British National Corpus a été créé par un consortium. Le BNC est un corpus annoté et balisé, donc avec une valeur ajoutée importante. Le BNC a fixé de nouvelles normes d'excellence dans la création de corpus, mais est également figé dans le temps, alors que le corpus COBUILD a continué d'évoluer, pour devenir l'actuel *Bank of English*.
Tandis que le corpus COBUILD était extrêmement important en taille pour son époque, d'autres corpus plus petits ont également été créés pour les besoins des études dans les langues de spécialité au sein de l'école de Birmingham.
Ce que nous appelons l'école de Birmingham a commencé dans les années soixante autour de Sinclair et Coulthard. L'école était concernée par les applications dans l'enseignement de la linguistique appliquée. Ainsi nous trouvons la tradition, personnifiée par Palmer, de la recherche appliquée. L'analyse de discours, surtout le discours scientifique, dans le but d'enseigner les langues de spécialité était centrale. Le texte de Barbier (1962) sur les caractéristiques des articles de recherche était le début des analyses sur le genre de Swales (1990). Tandis que Swales et d'autres travaillaient sur l'analyse des textes scientifiques, Roe (1977) travaillait sur un corpus scientifique jetant les bases pour les nombreuses études sur l'anglais de spécialité de l'Université d'Aston.

3. L'ère actuelle

La suite du développement de la linguistique de corpus est liée à la démocratisation des outils informatiques et des ressources électroniques. D'abord l'avènement des clones PC, en commencant avec l'Amstrad, et les Mac-Apple a rendu l'outil disponible à un plus grand nombre. En même temps nous avons vu l'arrivée des concordanciers comme Microconcord (Scott & Tribble) et ATA (Aston Text Analyser de Roe) pour DOS et Conc pour Mac. Il faut souligner que le but n'est pas le développement des outils, mais l'emploi des outils pour regarder les mots en contexte à travers le mot-clé en contexte, KWIC. En linguistique de corpus contextualiste, l'outil informatique n'est qu'une loupe pour mieux voir. L'intérêt se trouve dans le détail : pouvoir

généraliser est important, mais non pas formaliser. Ce que nous observons est un réseau de choix, suivant le principe d'idiome (Sinclair 1991). À ce stade, il n'y avait que deux moyens pour obtenir des données : les entrer manuellement, ou utiliser un scanner, un outil encore rare. Il est possible qu'à cette époque les critères de création de corpus aient été mieux suivis : quand les documents sont difficiles à obtenir, on fait plus attention au choix des textes.
L'avènement de Windows a encore simplifié les choses, d'autant plus qu'Internet est rapidement arrivé avec un choix de plus en plus important de documents. Les premiers concordanciers travaillaient uniquement sur du texte ASCII, pour traiter le html, puis le sgml : il a fallu faire évoluer les outils. Ainsi, Microconcord s'est mué en WordSmith Tools (Scott – www.lexically.net) et Conc en MonoConc (Barlow – www.athelstan.com), dorénavant disponible pour Windows. Puis, plus tard le BNC est devenu disponible sur CD-ROM, accompagné de SARA, qui est maintenant devenue XAIRA, outil pouvant traiter tout corpus en XML, même très basique.

3.1. Qui fait quoi ?

On peut distinguer cinq grands centres de linguistique de corpus, l'Université de Birmingham avec l'équipe de Sinclair, et maintenant Teubert, son successeur dans la chaire de Harper Collins, l'Université d'Aston à Birmingham avec Roe, l'Université de Liverpool autour de Hoey et Scott. Et puis il y a le centre de Lancaster, beaucoup plus TAL dans son approche fondée sur les travaux de Leech, et Oxford, maison mère de la TEI en Europe. Il y a évidement d'autres centres qui se créent avec le mouvement des chercheurs.
Les trois premiers restent plus contextualistes avec un minimum d'intervention sur le corpus, puisque Sinclair défend l'idée de zéro annotation (Sinclair 2005). Le but reste largement l'enseignement des langues, surtout les langues de spécialité, et le développement de la lexicographie. L'autre école se tourne vers des approches plus larges dans la création d'outils d'annotation et les applications typiquement TAL. Cependant, il ne faut pas une histoire de chapelles avec des écoles distinctes. Il y a simplement un continuum avec un glissement vers le TAL dans un sens, et vers d'autres disciplines de la linguistique appliquée dans l'autre.
La linguistique de corpus, *corpus linguistics*, s'est taillée une place de choix dans la linguistique appliquée. La meilleure introduction à l'approche contextualiste reste le livre de Sinclair (1991) 'Corpus, Concordance, Collocation'. La différence entre l'approche contextualiste inductive, *corpus-driven*, et d'autres méthodologies est décrite par Tognini-Bonelli, travaillant dans le cadre de l'école de Sinclair. Pour une introduction à la discipline, il faut lire Kennedy (1998), ou Hunston (2002) pour les applications en linguistique appliquée.

3.2. Que font-ils ?

La linguistique de corpus est une linguistique appliquée, la théorie est issue de la pratique, et non l'inverse. La langue est atteinte à travers la parole (Tognini-Bonelli 2001) et n'a pas d'existence propre en dehors du contexte. Ainsi, la linguistique de corpus se trouve en poursuivant la tradition établie par Palmer dans l'enseignement de l'anglais comme langue seconde à des non-spécialistes. Des études sur des corpus scientifiques visent à analyser des problèmes phraséologiques dans l'écrit scientifique (Gledhill 2000) ou la création de dictionnaires d'aide à la rédaction (Williams 2002a). Ces deux derniers étaient des étudiants de Roe, lui-même issu de l'école de Birmingham et élève de Sinclair. L'analyse des corpus scientifiques, soit comme étude linguistique, soit comme aide à la rédaction, est un thème récurrent dans la linguistique de corpus contextualiste (Tognini-Bonelli & Del Lungo Camiciotti 2005). Toujours dans l'enseignement, d'autres travaillent pour faire entrer le concordancier dans la salle de classe (Sinclair (éd.) 2004, Gavioli 2005).
Les applications de la linguistique de corpus sont nombreuses (Hunston 2002), et incluent la linguistique légiste, domaine développé par Coulthard (1994). D'autres études concernent la terminologie (Pearson, 1998) ou la traduction (Kenny, 2001).
Dans les domaines plus linguistiques, Hunston & Francis (2000) ont mené des études sur des grammaires locales utilisant le corpus COBUILD. Williams (1998, 2002b) a exploré les réseaux thématiques dans un corpus spécialisé et utilise la collocation comme outil de catégorisation. Les patrons thématiques et les mots-clés sont le sujet de nombreuses études (Scott & Tribble, 2006). L'analyse de discours sur corpus est un autre domaine important (Stubbs, 1996, Partington et al. 2004).
Cette liste est loin d'être exhaustive. Le paradigme contextualiste en linguistique de corpus est

employé partout dans le monde, sur l'anglais et d'autres langues. Je n'ai pas non plus parlé de l'autre grande tradition de linguistique de corpus représentée par l'ICAME. Les approches sont nombreuses, mais l'objet d'étude reste un corpus constitué selon des critères linguistiques (Sinclair 2005). L'objet est le corpus, les outils informatiques ne sont que des outils pour mieux voir dans le corpus, les objectifs sont toujours une meilleure compréhension du langage parlé par les êtres humains pour les êtres humains, c'est-à-dire la communication.

4. Conclusion

En guise de conclusion, il est temps de faire un petit rappel. Cette communication n'entre pas dans la rubrique histoire de la linguistique. Je ne retrace pas des origines pour faire de l'histoire, mais pour expliquer des paradigmes de recherche actuels. Ce n'est pas non plus pour prouver qu'un paradigme est meilleur qu'un autre, mais que les paradigmes existent, et qu'il faut les regarder et les comprendre afin de créer des échanges et d'avancer dans la recherche sur le sable mouvant que constitue le langage.

La linguistique de corpus est largement issue du monde anglo-saxon, et en anglais le mot linguistique est invariable, c'est une seule et unique discipline avec une multitude de facettes. Parmi ces facettes se trouve la linguistique de corpus : par le jeu de la collocation si chère à Firth, le mot corpus a pris un sens particulier. Il s'agit d'un ensemble de textes soigneusement choisis pour les besoins de la recherche linguistique et qui cherche à représenter une partie de la langue en action. Dans ce sens l'environnement de la langue, avec tous les aspects sociolinguistiques, doit être pris en compte. C'est-à-dire, le contexte culturel et le contexte situationnel. Pour un linguiste de corpus contextualiste il n'est nullement besoin de mettre ces paramètres dans une définition de corpus, c'est un acquis, cela va de soi depuis Malinowski. Dire que le sens du mot corpus est plus restreint en linguistique de corpus n'est pas dire qu'il ne peut pas y avoir d'autres types de corpus, simplement que l'association des mots linguistique et corpus a créé des attentes plus restreintes. Les autres corpus, juridique, littéraire existent, et on peut en faire des études linguistiques : ainsi il existe une linguistique *sur corpus* à côté de la linguistique *de corpus* où la constitution du corpus est en soi une partie essentielle de l'étude.

La ou les linguistiques, je ne vois pas la nécessité d'éclater une discipline sur une simple particule. Le TAL n'est pas la linguistique de corpus, la pragmatique ou la sociolinguistique non plus, chacune a son propre but. Cependant, ils peuvent utiliser les corpus, mais nous sommes de retour sur la linguistique de corpus.

Si la linguistique de corpus existe comme discipline autonome, où se trouvent les frontières avec d'autres disciplines ? Là, je retourne la question : avons-nous vraiment besoin de frontières quand toutes nos propres études sur le langage prouvent que les frontières n'existent pas ? La linguistique de corpus, comme d'autres disciplines de la linguistique, rentre parfaitement dans la notion de prototype, avec un nœud central et une périphérie qui glissera subtilement vers d'autres disciplines dans un continuum. Les catégories n'existent pas en soi, nous les créons pour mieux saisir la complexité. Parler des linguistiques de corpus est noyer le poisson, si tout le monde le fait, personne ne le fait, et tout le monde est perdant. La linguistique de corpus existe, elle est récente et sa méthodologie et son épistémologie se forgent. Pour la forger, il faut simplement la reconnaître.

BIBIOGRAPHIE

BARBER, C.L. 1962. Some Measurable Characteristics of Modern Scientific Prose, in J. Swales *Episodes in ESP*, Hemel Hempstead, Pergamon Press, pp. 3-14.

BAZELL, C. E, CATFORD, J. C., HALLIDAY, M. A. K., ROBINS, R. H. (éds.) 1966. *In Memory of JR FIRTH*, London, Longman.

COULTHARD, M. 1994. On the use of corpora in the analysis of forensic texts, *Forensic Linguistics*, 1, pp. 27-44.

FIRTH, J.R. 1935. *The Semantic of Linguistic Science*, in J.R. Firth, 1957. Papers in Linguistics 1934-1951, Oxford, OUP. 1948.

GAVIOLI, L. 2005. *Exploring corpora for ESP Learning*, Amsterdam, John Benjamins.

GLEDHILL, C. J. 2000. *Collocations in science writing*, Tübingen, Gunter Narr Verlag.

HALLIDAY, M. A. K. 1961. Categories of the Theory of Grammar, *Word*. 17.3, pp. 241-92.

HALLIDAY, M. A. K. 1966. Lexis as a linguistic level, in C. E. Bazell et *al.*, pp. 148-162.

HALLIDAY, M.A.K., HASAN, R. 1976. *Cohesion in English*, London, Longman.
HOEY, M. 1991. *Patterns of Lexis in Text*, Oxford, Oxford University Press.
HOEY, M. 2005. *Lexical Priming: A New Theory of Words and Language*, London, Routledge.
HOWATT, A.P.R. 1984. *A History of English Language Teaching*, Oxford, OUP.
HUNSTON, S., FRANCIS, G. 2000. *Pattern Grammar: A corpus-driven approach to the Lexical Grammar of English*, Amsterdam et Philadelphie, John Benjamins.
HUNSTON, S. 2002. *Corpora in Applied Linguistics*, Cambridge, CUP.
KENNEDY, G.1998. *An introduction to corpus linguistics,* London & New York, Longman.
KENNY, D. 2001. *Lexis and Creativity in Translation*, Manchester, St Jerome Publishing.
LÉON, J. 2005. Claimed and unclaimed sources of *Corpus Linguistics, Henry Sweet Society Bulletin*, N°44, pp. 36-50.
MALINOWSKI, B. 1923. The problem of meaning in primitive languages. Supplement to CK. Ogden and I.A. Richards, 1923. pp. 296-336.
MALINOWSKI, B. 1935. Coral Islands and their Magic, vol 2. The language of Magic and gardening, London, George Allen and Unwin Ltd.
OGDEN, C.K., RICHARDS, I.A. 1923. *The Meaning of Meaning*, London, Routledge and Kegan Paul.
PALMER, H. E. 1933. *Second Interim Report on English Collocations*, Tokyo, Kaitakusha.
PARTINGTON, A., MORLEY, J., HAARMAN, L. (éds) 2004. *Corpora and Discourse : Proceedings of CamConf 2002 Università degli Studi di Camerino, Centro Linguistico d'Ateneo Sept 27th-29th 2002.* Bern, Berlin, Bruxelles, Frankfurt/M., New York, Oxford, Wien, Peter Lang.
PEARSON, J. 1998. *Terms in Context*, John Benjamins.
ROE P. 1977. *Scientific Text*, ELR University of Birmingham.
SINCLAIR, J. McH. 2005. Corpus and Text: Basic Principles, in M. Wynne (éd.). 2005. pp. 1-16.
SINCLAIR, J. McH. 1991. *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.
SINCLAIR, J. McH., JONES, S., DALEY, R. 2004. *English Collocation Studies: The OSTI Report*, Londres - New York, Continuum.
SINCLAIR, J. McH. (éd.) 1987. *Looking Up: an account of the COBUILD Project in Lexical Computing*, London, Collins.
SINCLAIR, J. McH. (éd.) 2004. How to use corpora in language teaching, Amsterdam, John Benjamins.
SINCLAIR, J. McH. 1966. Beginning the study of lexis, in C. E. Bazell et *al.* 1966. pp. 410-430.
SINCLAIR, J. McH. et al., 1970. *English Lexical Studies: Report to OSTI on Project C/LP/08,* Department of English, University of Birmingham.
SWALES, J. M. 1990. *Genre Analysis*, Cambridge, Cambridge University Press.
TOGNINI-BONELLI, E., DEL LUNGO CAMICIOTTI, G. (éds.) 2005. *Strategies in academic discourse*, Amsterdam, John Benjamins.
TOGNINI-BONELLI, E. 2001. *Corpus Linguistics at Work*, Amsterdam, John Benjamins.
TUTIN, A., GROSSMAN, F. 2003. *Les collocations : analyse et traitement*, Amsterdam, de Werelt.
WILLIAMS, G. 1998. Collocational Networks : Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles, *International Journal of Corpus Linguistics*, Vol. 3/1, pp. 151-171.
WILLIAMS, G. 2002a. Corpus-driven lexicography and the specialised dictionary: headword extraction for the Parasitic Plant Research Dictionary, in A. Braasch, C. Povlsen (éds), 2002, *Proceedings of the 10th EURALEX International Congress*, Copenhague, CSK, pp. 859-864.
WILLIAMS, G. 2002b. In search of representativity in specialised corpora: categorisation through collocation, *International Journal of Corpus Linguistics*, Vol. 7/1, pp. 43-64.
WILLIAMS, G. 2003. Les collocations et l'école contextualiste britannique, in A. Tutin et F. Grossman, *Les collocations : analyse et traitement*, Amsterdam, de Werelt, pp. 33-44.
WYNNE, M (éd). 2005. *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford, AHDS

(EN)-JEUX DE CORPUS POUR LA RECHERCHE EN SHS. ÉNONCÉS, TEXTES ET DOCUMENTS

Huguette RIGOT
MCF SIC Paris X / INRP

SOMMAIRE

Résumé : *La notion de corpus est à caractériser, connotant des réalités et des objets textuels différents suivant les disciplines et les situations de recherche. Pour ce faire, les pratiques issues de trois traditions savantes sont utiles à analyser.*
La linguistique de corpus fait figure d'exemple dans les SHS. La réflexion de la linguistique de corpus a permis non seulement de spécifier cette notion, mais aussi de développer de nouveaux champs disciplinaires souvent, d'ailleurs articulés à d'autres disciplines comme l'histoire, l'analyse de discours en est certainement le meilleur exemple.
Mais s'il existe aujourd'hui un dynamique ensemble de réflexions portant sur la notion de corpus, certains secteurs de la recherche n'ont pas encore intégré les potentialités des notions de corpus et de numérique. Quand pourra-t-on parler d'une sociologie du corpus ? Cette question que se posent aujourd'hui chercheurs et institutions est particulièrement pertinente et y répondre est urgent. La constitution de corpus de données d'enquêtes quantitatives et surtout qualitatives correspond à plusieurs objectifs qui convergent tous vers un repositionnement épistémologique et réflexif des sciences de la culture.

Introduction

Le chrononyme « société de l'information » nous projette dans l'ère du numérique. Il transforme de manière radicale notre rapport à l'écrit et notre rapport à l'autre. Les changements qui, actuellement, nous affectent, ont une dimension à la fois sociale, cognitive et sémiotique.
Pourtant, assez paradoxalement, la plupart des pratiques de recherche en sciences humaines et sociales ne sont concernées qu'à la marge par l'évolution ou la révolution apportée par le numérique.
Aussi, il semble intéressant d'analyser en premier lieu quelles sont les possibilités ouvertes par les nouvelles technologies de l'information, notamment par la constitution de corpus numériques et/ou de bases de données textuelles regroupant des matériaux issus d'enquêtes qualitatives, puis dans un deuxième temps de confronter ces possibilités à l'existant, par l'analyse du traitement actuel appliqué à ces données d'enquête par leurs producteurs et enfin de considérer comment le positionnement épistémologique des sciences humaines et sociales peut évoluer grâce à la constitution de ces corpus numériques.
La lecture de différents rapports sur le statut des données d'enquêtes qualitatives permet, à partir de la reconnaissance ou de la non-reconnaissance de la valeur des enquêtes qualitatives par diverses communautés académiques, de constater que ce problème est abordé de deux manières, soit par le biais de la langue donc des corpus oraux et soit par celui du recueil des données, focalisé sur la pratique d'entretien générant des corpus de données orales à la fois nombreux et volumineux.

Notre hypothèse de travail est que ces deux voies sont complémentaires. Pourtant, en prenant les disciplines qui les symbolisent le mieux, à savoir d'un côté la linguistique de corpus et de l'autre la sociologie, celles-ci travaillent encore peu de concert : les objectifs de connaissance poursuivis et surtout les modalités méthodologiques sont différents. La linguistique de corpus est une discipline d'observation des pratiques langagières, alors que la sociologie et toutes les sciences sociales et humaines, utilisant comme méthode de recueil de données l'entretien, l'immersion participante, etc., sont des disciplines d'interactions, possédant une forte valeur communicationnelle.
Ainsi, d'un côté l'observation, de l'autre les interactions ! Cette opposition prenant en compte des modes d'approche spécifique des réalisations langagières permet de comprendre d'une part pourquoi les données qualitatives sont à ce point « abandonnées », en déshérence et servent d'argument à la dévalorisation des résultats obtenus par les sciences humaines et sociales quand ceux-ci ne doivent rien aux statistiques et d'autre part comment se donner les moyens de modifier le statut épistémologique des sciences humaines et sociales en instaurant de nouvelles approches méthodologiques du traitement des paroles recueillies auprès de la société civile. Ces modes d'approche des réalisations langagières, en rendant compte d'un engagement différencié des chercheurs par rapport à ces données – observateurs des corpus oraux ou acteurs participant à des interactions avec des enquêtés -- sont des facteurs explicatifs à la fois du statut méthodologique accordé aux corpus oraux et des traitements qui leur sont appliqués.

1. Les corpus numériques et / ou bases de données textuelles : pourquoi et comment les constituer ?

1.1. Le corpus : une notion désormais centrale

En évoluant de l'introspection au corpus[1], aujourd'hui, la linguistique se définit principalement comme une discipline d'observation des choix linguistiques effectués par des locuteurs dans des contextes réels. Par ce changement, quatre éléments possibles émergent des analyses linguistiques :

- le corpus remplace le texte, qui lui-même a été l'objet d'une sorte de révolution en devenant une unité d'analyse plus complexe que le mot et la phrase,
- les réalisations langagières sont produites par des acteurs réels, ordinaires et ou appartenant à des communautés spécifiques. La variation linguistique est devenue ainsi un objet d'analyse, tout en s'appuyant sur des faits de langue authentiques.
- les possibilités d'accéder à des réalisations langagières sont devenues infinies et bien évidemment, le web peut être aujourd'hui d'une part considéré comme un réservoir illimité à la fois par le statut, la variation des réalisations présentes et surtout par leur volume et d'autre part comme le moyen d'accéder à des bases de données textuelles et des corpus numériques constitués spécifiquement pour l'analyse linguistique, la base de données Frantext en est l'exemple le plus significatif.
- la comparaison des pratiques langagières est désormais devenue non seulement possible, mais un des fondements de cette nouvelle linguistique.

1.2. Qu'est-ce qu'un corpus ?

Si cette notion de corpus linguistique est devenue si centrale, elle nécessite d'être définie et caractérisée. François Rastier, sous forme métaphorique, parle de « sac de mots ou archives de textes[2] », évidemment pour proposer une caractérisation plus scientifique.
La notion de corpus renvoie traditionnellement à deux conceptions. La première est documentaire et ne retient que des variables globales ignorant les aspects textuel et structurel. Dans ce cas, le corpus est un réservoir d'exemples langagiers ou … une base de données textuelles.
La deuxième conception, plus liée à une tradition herméneutique, prend en compte les relations intertextuelles.

---

[1] Pour s'orienter efficacement sur quelques textes fondamentaux et récents sur la linguistique de corpus, on peut citer : *La linguistique de corpus* sous la dir. de Geoffroy Williams, Presses universitaires de Rennes, 2005 et la revue *Corpus.* Numéro 1 « Corpus et recherches linguistiques », novembre 2002. Notamment les articles de Jean-Philippe Dalbera, « Le Corpus entre données, analyse et théorie » et Damon Mayaffre, « Les corpus réflexifs : entre architextualité et hypertextualité ».

[2] François Rastier, « Enjeux épistémologiques de la linguistique de corpus » dans *La Linguistique de corpus* sous la dir. de Geoffroy Williams, Presses universitaires de Rennes, 2005, p. 31.

La définition proposée par F. Rastier fait du « corpus (…) un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications ».[1]
Ainsi, tout regroupement de textes ne peut être considéré comme un corpus… il peut être simplement une base de données textuelles.
Un corpus suppose une préconception, c'est-à-dire une explicitation du choix des textes qui le composent et donc une sélection. Il implique la définition d'un objectif d'usage, donc d'un projet scientifique. Cette caractéristique est fondamentale, puisqu'elle permet de dire que le corpus n'est pas représentatif de la langue, mais qu'il est un **construit**[2] adéquat à un objet ou à une tâche qui détermine alors ses critères de représentativité et son degré d'homogénéité. Dans tous les cas, le corpus dépend du point de vue académique et aussi « personnel » du chercheur. Le corpus doit être à la fois construit et « aimé » par celui qui le rassemble et l'analyse.
S'il est un construit, le corpus est situé dans des pratiques qui « travaillent », documentent, catégorisent des rassemblements textuels, ainsi, quatre niveaux sont à distinguer : l'archive qui regroupe l'ensemble des documents accessibles, le corpus de référence à partir duquel des corpus d'études vont être délimités et enfin le sous corpus de travail variant selon les étapes de l'analyse.
Ainsi, la notion de corpus linguistique étant stabilisée et opératoire, de quelle façon peut-elle être utilisée pour organiser des rassemblements de textes en sciences humaines et sociales ? La réflexion, qui a conduit à constituer la linguistique de corpus peut-elle aider à faire comprendre l'importance, du point de vue de la linguistique et plus généralement du point de vue des sciences humaines et sociales, du rassemblement, donc de la préservation en vue de la communication des matériaux langagiers produits par les enquêtes qualitatives ?

2. Préserver pour communiquer : une nouvelle approche des matériaux langagiers en sciences humaines et sociales

Il s'agit bien évidemment des données qualitatives, entretiens enregistrés, notes écrites lors d'observation participante, etc.
De ce fait, il faut écarter les données quantitatives. Pourtant, en France leur traitement n'est peut-être pas sans rapport avec celui dévolu aux données qualitatives.
Leur situation a été récemment et peut-être momentanément réglée à partir du Rapport[3] de Roxane Silberman du Lasmas - CNRS qui constatait que les données publiques produites avec des financements publics n'étaient pas accessibles directement aux chercheurs individuels et aux laboratoires, les grands producteurs de données quantitatives comme l'INSEE faisant payer leur utilisation. Ainsi est né le centre Quetelet[4] dont la fonction est d'archiver les données des différents producteurs de la Statistique publique à des fins de diffusion et de réutilisation pour les chercheurs. Cette utilisation fait l'objet d'une réglementation. Cet accès est limité aux fonctions de recherche ou d'enseignement, pour écarter les réutilisations purement commerciales et il se fait selon une réglementation et à partir d'un engagement formel et écrit des utilisateurs. Le problème de la préservation pour communication des données quantitatives semble réglé.
On peut raisonnablement penser que cette nouvelle modalité d'accès permet de dynamiser les recherches. La plupart des laboratoires, avant la création du Centre Quetelet, se contentaient de citer des données sous forme de schémas, tableaux, etc. repris de publications précédentes. Ainsi, le lecteur averti lisait plusieurs fois les mêmes informations sans avoir vraiment l'impression d'une réutilisation critique. Mais pour critiquer les sources… faut-il encore y avoir accès.

2.1. La variété et le statut des matériaux langagiers

Que ce soit dans les rapports Français ou étrangers, quand on évoque ces corpus numériques de données qualitatives, il est quelquefois difficile de savoir de quel type de données on parle. Nous

---

[1] *Op. cit.* p. 32
[2] Marie-Paule Jacques, « Pourquoi une linguistique de corpus ? » dans *La Linguistique de corpus* sous la dir. de Geoffroy Williams, Presses universitaires de Rennes, 2005, p. 26.
[3] Consultable à l'adresse http://www.ladocumentationfrancaise.fr/rapports-ublics/004000935/0000.htm (consulté 28 juin2006)
[4] Consultable à http://www.centre.quetelet.cnrs.fr/

sommes dons assez loin de la caractérisation précise de la notion de corpus telle qu'elle est présente dans la linguistique. De façon assez large, il peut s'agir de trois types de données :
a) des revues en lignes et des archives ouvertes qui mettent à la disposition de lecteurs sur le net des résultats d'enquête publiés ou finalisés auxquels on accède par abonnement ou librement.

b) le plus souvent, il s'agit de données d'enquête non publiées : la « littérature grise » et qui pose là un véritable problème de préservation et de communication tant auprès des chercheurs qu'auprès d'un public plus élargi. À cet égard, le *Rapport Canadien de mai 2001 sur l'évaluation des besoins sur la Consultation nationale sur les archives de résultats de recherche*[1] est tout à fait révélateur de l'ambiguïté ou de la difficulté à définir et à caractériser l'objet de la conservation : littérature grise uniquement ou matériaux d'enquêtes qualitatives ?

c) d'autres rapports indiquent clairement que ce sont les matériaux des enquêtes qualitatives qu'il faut préserver en vue d'une communication. Même si le mode d'accessibilité souhaité est le net, l'état des données est à ce point préoccupant que le simple repérage d'ensembles documentaires, consultables dans des lieux publics, comme les dépôts d'archives, correspondrait déjà à une révolution dans le traitement des matériaux qualitatifs.

2.2. La déshérence des données, une raison de la dévalorisation des enquêtes qualitatives, le cas français

En avril 2003, Françoise Cribier, avec la collaboration d'Elise Feller, a rédigé et présenté au Ministère délégué à la Recherche et aux nouvelles technologies un rapport portant sur la conservation des données qualitatives des sciences sociales en France.[2] La rédaction de ce rapport se situe dans un contexte d'interrogation sur le statut et surtout le devenir des données issues des enquêtes qualitatives, c'est-à-dire des données orales. En l'espace de moins de dix ans plusieurs rapports ont été commandés par différentes institutions[3] et rédigés par des spécialistes se positionnant différemment, selon les commanditaires. C'est d'abord sous l'angle des archives sonores de diction -- une place importante a été faite aux archives radiophoniques et télévisuelles -- puis sous l'angle de l'archivage d'entretiens faits à l'occasion d'enquêtes qualitatives et enfin, sous l'angle de la conservation des fonds sonores que cette problématique a été abordée. Ainsi, trois préoccupations s'articulent ou s'entremêlent : les traces laissées par les pratiques langagières, les données d'enquêtes qualitatives et la préservation des documents sonores. Indiquer ce contexte n'est pas inutile pour comprendre comment une synergie se met en place ayant pour objectif de traiter, pour ce qui nous concerne ici, des données d'enquêtes qualitatives. Un entretien, par exemple, effectué dans le cadre d'une enquête, ressort bien à la fois de l'usage de la langue, de l'information scientifique qu'il a contribué à produire et enfin du support sur lequel il a été enregistré. Ce support est double. Il peut s'agir du support matériel, c'est-à-dire de bandes audio et/ou audiovisuelles, mais aussi du support des transcriptions transformant les entretiens d'énoncés oraux en textes écrits qui, eux aussi, sont liées à des types de supports différents. Une transcription d'entretien, sans aborder tout de suite le problème du codage ou du « nettoyage » des données, suppose sa transformation en texte écrit sur du papier ou en fichier informatique.

---

[1] Consultable à http://www.sshrc.ca/web/about/publications/da_phase1_f.pdf (consultée 28 juin 2006)

[2] Françoise Cribier, *Projet de conservation des données qualitatives des sciences sociales recueillies en France auprès de la « société civile ».* Rapport présenté au Ministère délégué à la Recherche et aux nouvelles technologies en avril 2003.
http://www.iresco.fr/labos/lasmas/rapport/Rapdonneesqualita.pdf, consultée le 15 juin 2006.

[3] Les principaux rapports sont :
-- le rapport remis au Conseil économique et social en janvier 2001 par Georgette Elgey avec la collaboration d'Annette Wieworka portant sur l'ensemble des archives sonores de diction
-- le rapport, sous la direction de Claude Dubar et du Laboratoire Printemps en 2001 pour le secteur SHS du CNRS, s'intéressant surtout aux entretiens d'enquête
-- le rapport rédigé par Marie-France Calas en 2001 pour le Ministère de la culture et de la communication aborde les problèmes de politiques de conservation et de valorisation menées en France pour l'ensemble des fonds sonores.
Il convient aussi de citer le *Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux*, sous la direction d'Olivier Baude et commandé par la Délégation générale à la langue française et aux langues de France. Ce guide vient d'être édité sous le nom de *Guide corpus oraux* par les éditions du CNRS.

Autre remarque, les disciplines n'ont pas toutes le même rapport à la parole de l'interviewé, psychologues et ethnologues savent mieux conserver la parole d'autrui. Leur formation les conduit d'une part à accorder une plus grande importance à ce type de matériau qui est souvent issu de situations singulières – l'entretien clinique, confidentiel pour le psychologue, l'observation, la participation, les rencontres exotiques pour l'ethnologue – et d'autre part à savoir comment matériellement les documenter pour les retravailler plus tard ou collectivement. Peu de sociologues ont ce souci de préservation et de partage.
En France, la situation des données d'enquêtes qualitatives est très préoccupante. Comment spécifier ce « contexte alarmant de l'archivage des sources orales »[1] ? Pour l'heure, aucune institution, aucune communauté -- sauf peut-être celle des ethnologues -- très peu de laboratoires et enfin de très rares chercheurs ont pour préoccupation de sauvegarder leur données d'enquête.
Au cœur du problème soulevé par la préservation et la sauvegarde de l'ensemble des données d'enquête, se situent la comparaison, la cumulativité des résultats, c'est-à-dire l'inscription des recherches dans une articulation de temporalités, celle de la recherche elle-même, celle des effets produits et enfin celle du regard distancié par le temps, les changements sociaux et les paradigmes scientifiques, qui sont directement en cause. Dans la majeure partie des cas, les résultats d'enquêtes, même quand ils sont publiés, sont accessibles au grand public et au public de chercheurs sous un format particulier, c'est-à-dire un nombre de caractères, de pages, un appareil informationnel plus ou moins développé, imposé par un éditeur à partir de contraintes commerciales et non scientifiques.
Données perdues, égarées, rangées dans les placards, détruites par manque de place ou à la suite de déménagements, donnent une image de la recherche et de l'importance de la méthodologie tout à fait contraire à celle qui prévaut dans les ouvrages de méthodologie d'enquête. De plus, le temps passé au recueil, au traitement, à la documentation de ces données est bien plus important que le temps passé à la rédaction des résultats de recherches. Pourtant ce sont ceux-ci qui restent, s'ils sont édités, qui servent à produire des interprétations, ce sont eux qui témoignent de la science-en-train-de-se faire, ce sont eux qui finissent par représenter la science faite[2]., Ce sont ces résultats de recherches qui servent à évaluer et la qualité de l'enquête et la qualité des chercheurs. À ce propos, il faut remarquer que, particulièrement dans le rapport rédigé par Françoise Cribier, la qualité de l'écriture des résultats d'enquêtes a orienté l'auteur vers certains chercheurs à interviewer, établissant ainsi une hypothèse implicite, peut-être à vérifier, à savoir que de bons travaux de recherche, des résultats bien rédigés et publiés sont sous-tendus par de « bonnes pratiques » de recueil et de traitement des données.
Force est de constater une fracture entre ce qui est dit, écrit dans les manuels de recherche à l'usage des jeunes chercheurs, -- le terrain et les entretiens sont des actes quasi initiatiques -- et ce que les chercheurs font réellement de leurs données, une fois les résultats écrits et publiés, c'est-à-dire un abandon, une mise au placard, un rejet. Mais, peut-il en être autrement en l'absence d'une formation à la sauvegarde des données, en l'absence de centres d'archivage et peut-être surtout en l'absence d'une conscience que ces données appartiennent aux chercheurs, aux enquêtés et aux institutions commanditaires, en l'absence d'une conscience de la valeur patrimoniale des paroles recueillies auprès de la société civile.

2.3. L'engagement du chercheur dans ses données

Si la situation d'un point de vue général est particulièrement alarmante, certaines communautés scientifiques, comme les ethnologues, et certains professionnels -- archivistes, documentalistes -- certainement sous la double impulsion donnée d'une part par la linguistique de corpus et d'autre part par le développement des nouvelles technologies, permettant la numérisation de grandes masses de documents et surtout leur mise à disposition, ont décidé de réagir et de mettre à profit leur savoir-faire professionnel et scientifique pour stopper la perte systématique des données qualitatives, et surtout pour renouveler en profondeur le travail scientifique.
Le désintérêt pour ces données qualitatives porte témoignage d'une double conception de ce qu'est faire de la recherche en sciences humaines et sociales.
En premier lieu, il faut s'interroger sur la raison de la déshérence de ces données et comprendre que cette situation a parti lié avec le statut épistémologique des sciences humaines et sociales.

[1] Rapport de Claude Dubar (ce rapport n'ayant jamais fait l'objet d'une large communication, n'est pas accessible).
[2] En référence aux travaux de sociologie des sciences, notamment ceux de Bruno Latour.

Les matériaux qualitatifs sont considérés comme peu dignes de confiance, ils témoignent d'un certain type de vérité, celle rapportée par les enquêtés, mais qui n'est pas celle des scientifiques. Travailler à partir de tels matériaux, c'est s'exposer à un certain nombre de biais scientifiques, quasiment tous répertoriés. D'abord ce sont des sources provoquées donc biaisées du fait des circonstances de leur élaboration. Ensuite, ils sont le produit de ce que Pierre Bourdieu appelait l'illusion biographique[1]. Il distinguait ainsi trois définitions de l'illusion : l'illusion téléologique surestimant l'intentionnalité car recomposant après-coup des événements rassemblés pour atteindre un objectif, l'illusion de rester soi-même qui permet à l'individu de récupérer son unité à travers la complexité des situations vécues, et enfin l'illusion de personnalité permettant à l'individu de se sentir différent des autres.
En réponse à ces objections qu'elle considère comme étant réelles, Françoise Cribier, préfère admettre que ces trois grandes illusions correspondent aux réalités complexes de nos sociétés.
De plus, l'utilisation de ces matériaux qualitatifs demande à la fois une documentation sérieuse et une critique vigilante. Mais n'est-ce pas le travail même des scientifiques que de vérifier et de critiquer leurs sources ? Dans les recherches où ce travail n'a pas été mené, l'utilisation des matériaux qualitatifs est différente : au lieu d'étayer les éléments théoriques, ils servent d'illustration ou de confirmation aux résultats obtenus par d'autres sources. Ils sont donc détournés de ce qui fait leur spécificité et leur valeur : appréhender « les réalités » des acteurs ordinaires de la société civile.
En deuxième lieu et raison suprême à notre sens, la personne, la parole du chercheur tout comme celle de l'interviewé sont dans les matériaux qualitatifs. La situation d'interaction créée lors d'un entretien révèle autant sur l'un que sur l'autre et peut-être encore plus sur le chercheur lui-même, surtout quand il a « raté » son entretien, quand il n'a pas su entendre et lire l'importance de ce qui lui était transmis, quand il a surexploité ses données pour confirmer son hypothèse. *Les données, ça ne se partage pas, d'ailleurs çà n'aurait aucun sens,* voilà le credo de la plupart des chercheurs *et puis si çà tombait dans des mains peu amicales, des chercheurs concurrents pourraient « repomper » sans citer leurs sources, c'est ce qui est d'ailleurs arrivé à Bourdieu dans La Misère du monde, des chercheurs qui pourraient mal interpréter et ainsi alimenter des querelles stériles et puis qui va conserver tout cela, quelles institutions peuvent garantir la préservation, la communication et surtout la pérennité du travail considéré à juste titre comme bien plus important à cause de la documentation accompagnant les matériaux ?* Toutes ces questions sont posées dans les rapports. Elles témoignent de la même méfiance, de la même préoccupation : communiquer ce qui est personnel, sans l'avoir mis à distance par exemple par la mise en place de protocole de documentation, est difficilement acceptable. Les chercheurs préfèrent abandonner leurs données plutôt que de les partager, de les communiquer. Ils préfèrent les abandonner plutôt que de devoir les retravailler pour les transformer en corpus numérique.

2.4. Les modalités de constitution des corpus numériques

Je n'en évoquerai que les grandes lignes.
Premier principe : il est moins coûteux de prévoir et de préparer la constitution de ces corpus avant et pendant l'enquête que par la suite. Documenter une enquête est un travail habituel, mais écrire et catégoriser donc normaliser cette documentation est autre chose. Seuls les chercheurs ayant appris à le faire dans leur formation et ayant l'habitude de travailler collectivement et donc de partager leurs données savent le faire. Documenter, cela consiste entre autre, à spécifier le contexte général de l'étude et chacune des situations d'enquête créée, comme les entretiens. Cela consiste aussi à établir des connexions avec d'autres situations, à garder la trace systématique des impressions, des événements particuliers qui peuvent sur le moment ou après permettre d'approfondir et de réorienter certaines interprétations.
Deuxième principe : respecter les dispositifs juridiques, donc faire signer un accord aux interviewés pour être questionnés et pour que leur parole soit communiquée en partie ou intégralement. Enfin, garantir l'anonymat.
Troisième point, le plus difficile à mettre en place, c'est-à-dire le traitement des données : les problèmes techniques relatifs à la prise de son et à l'enregistrement de terrain, puis les problèmes relatifs aux transferts sur des supports durables et enfin, les problèmes plus « intellectuels » de la transcription, du codage, du nettoyage des données et enfin de l'analyse de la qualité des données qui engagent directement le quatrième point ….

[1] Pierre Bourdieu, « L'illusion biographique » dans les *Actes de la recherche en sciences sociales*, 1986

Quatrième point : Que sélectionner ?
Cinquième point : Comment constituer des bases de données textuelles… Quelle plateforme logicielle ? Quel financement ? Quelle garantie institutionnelle ?
Sixième et dernier point qui relève à la fois de la maîtrise et des choix des chercheurs et des formats de données numériques : quels logiciels d'analyse des données ?

3. L'impact sur le statut épistémologique des sciences humaines et sociales

Le regard porté sur les méthodologies qualitatives est souvent ambivalent : d'un côté, une reconnaissance de richesse et de l'autre une méfiance (peu de généralisation, peu ou pas de contrôle) Pourtant, nombre de chercheurs, sociologues (cf. *Enquête de terrain*, dir. Par D. Céfaï[1]) spécialistes des SIC (cf. *Dictionnaire des recherches qualitatives en sciences humaines*, dir. A. Mucchielli [2]) ont travaillé à élaborer des recueils de méthodologies qualitatives pour chercher à comprendre les phénomènes sociaux à travers les réalisations langagières des individus.
Notre propos est de considérer comment, dans cette filiation d'élaborations méthodologiques, la constitution de corpus de données d'enquête et leur mise en accès sur le réseau devraient permettre d'ouvrir les méthodes qualitatives à de nouvelles perspectives, d'une part, pour leur donner une place reconnue et acceptée par les acteurs des sociétés civiles et politique et par ceux de la communauté scientifique et d'autre part pour améliorer leurs « performances » épistémologiques par la pratique de la cumulativité et de la comparabilité des données, permettant ainsi de produire des méta-analyses.
Ainsi, on peut faire l'hypothèse que la déshérence des matériaux qualitatifs a contribué à disqualifier les recherches qualitatives qui sont, de par leur objectif et de par leur méthodologie, des modalités d'approche du social, uniques et spécifiques. La maltraitance des données d'enquêtes qualitatives portant à la fois sur l'impréparation du traitement et sur leur abandon, une fois publiés les résultats, a un impact direct sur le développement des sciences humaines et sociales. Faire ce constat, c'est donc se demander pourquoi changer, pourquoi transformer les pratiques, pourquoi conserver ces données.
La **première raison est d'ordre patrimonial** : les paroles recueillies dans des contextes particuliers d'enquête ne pourront jamais l'être de nouveau. C'est un matériau qui périt avec celui qui l'exprime et qui disparaît une fois l'interaction entre l'enquêteur et l'enquêté interrompue.
La **deuxième raison est la réutilisation de ces matériaux**. La lecture par un autre chercheur au moment de l'enquête ou plus tard, c'est-à-dire un autre regard social et culturel porté sur ce qui a été dit et observé, peut engendrer une autre interprétation. De plus, des comparaisons avec des travaux de la même époque ou des travaux d'époques différentes peuvent être établies. C'est la cumulativité des matériaux et des résultats de recherche et leur insertion dans une temporalité qui permettent une évolution significative de la recherche en sciences humaines et sociales.
La **troisième raison est d'ordre méthodologique**. Ce retour sur les matériaux implique aussi un regard rétrospectif sur les méthodes utilisées par les chercheurs et sur les conditions de production de leurs travaux. Une histoire des disciplines ou des champs disciplinaires serait ainsi possible. De plus, des chercheurs d'autres disciplines, en accédant à ces données « nouvelles » pour eux, pourraient ainsi développer la pluridisciplinarité et montrer la capacité des sciences sociales à explorer ensemble des objets communs, à fédérer les savoir-faire, à capitaliser les résultats, pour les revisiter, les réinterpréter et démultiplier les interprétations.
La **quatrième raison** considère que le regard porté sur l'acteur social n'est pas uniquement de l'ordre de l'objet d'étude : il se situe aussi à l'intérieur du processus de recherche et il est une figure de **l'engagement social et scientifique** que se doit d'avoir le chercheur.
La cinquième raison porte sur la complémentarité entre qualitatifs et quantitatifs. S'il y a peu de temps, les milieux académiques ont réagi pour assurer une conservation et une accessibilité aux données quantitatives, le même travail doit être accompli pour les données qualitatives, car de nombreuses recherches mêlent les deux approches.

Aborder le traitement des matériaux qualitatifs, c'est toucher réellement la méthodologie des enquêtes qualitatives. Transformer les pratiques d'enquêtes, la façon d'évaluer leurs procédures et leurs résultats, devient possible.

[1] Daniel Céfaï , l'Enquête de terrain, textes réunis, présentés et commentés, La Découverte, 2003.
[2] Dictionnaire des méthodes qualitatives en sciences humaines sous la dir. d'Alex Mucchielli, A. Colin, 2004.

Reconnaître l'aspect communicationnel[1] de la recherche qualitative, c'est intégrer l'engagement du chercheur dans le processus d'enquête, accepter le partage des données qualitatives et démultiplier l'interprétation. Ainsi, les sciences humaines et sociales peuvent sortir de leur situation de disciplines dominées.

La linguistique de corpus pointe la nécessité de rassembler des matériaux langagiers réels, de les documenter pour décrire leurs conditions de production et rendre possible leur communication et d'effectuer des traitements interprétatifs. Elle met à disposition des chercheurs des sciences de la culture une méthodologie, une réflexion conceptuelle et surtout des modes opératoires rendant compte de la valeur spécifique de ces données. Elle inscrit celles-ci simultanément dans une temporalité, un cadre de pensée et des modalités de traitement assistés ou non par ordinateur. La notion de corpus absorbée, réappropriée et adaptée par les chercheurs peut sortir les recherches qualitatives de leur ghetto épistémologique.

BIBLIOGRAPHIE

CEFAÏ, D. 2003. *L'Enquête de terrain,* Paris, La Découverte.

Corpus. Numéro 1 « Corpus et recherches linguistiques », novembre 2002.

CRIBIER, F. 2003. *Projet de conservation des données qualitatives des sciences sociales en France auprès de la « société civile »*. En ligne sur le site du Lasmas (CNRS-EHESS).

Dictionnaire des méthodes qualitatives 2004. *en sciences humaines,* Paris, A. Colin.

RIGOT, H. 2005. Crises communicationnelles dans le processus de recherche en sciences humaines et sociales, in M. Gabay, *Communiquer dans un monde en crise*, L'Harmattan.

[1] Huguette Rigot, « Crises communicationnelles dans le processus de recherche en sciences humaines et sociales » dans Michel Gabay, *Communiquer dans un monde en crise*, L'Harmattan, 2005.

MÉTHODOLOGIE TRANSDISCIPLINAIRE DE GESTION DE CORPUS POUR LES DISCIPLINES DE L'INTERACTION : RECHERCHE DE PRINCIPES DIRECTEURS

Hassan ATIFI, Christophe LEJEUNE, Goritsa NINOVA, Manuel ZACKLAD
Université de Troyes, Institut Charles Delaunay, Tech-CICO (FRE CNRS 2848)

SOMMAIRE

Introduction

Le laboratoire Tech-CICO est engagé depuis plusieurs années dans un travail inter– et transdisciplinaire qui associe des chercheurs en sciences humaines et en informatique, en particulier, dans le cas de recherches intervention auprès d'organisations professionnelles. Dans ce contexte, la nécessité de partager des corpus numérisés d'interactions communicatives et leurs commentaires entre différents analystes est de plus en plus manifeste. Pour surmonter les difficultés liées aux choix de corpus, aux modalités de segmentation et aux catégories d'analyse, nous avons décidé de lancer un projet de recherche interne visant à développer une méthodologie transdisciplinaire de gestion du corpus pour les disciplines de l'interaction (sélection des situations de référence, acquisition, retranscription, documentarisation, segmentation, interprétation, valorisation, etc.).

Notre objectif est de partir des pratiques concrètes de constitution de corpus dans plusieurs disciplines afin de parvenir à définir les conditions du partage matériel des sources relevant d'un intérêt commun et à expliciter les catégories d'analyse élaborées dans les différentes disciplines pour permettre un dialogue avec ses pairs. Pour mener cette recherche exploratoire nous avons procédé à la réalisation d'entretiens approfondis avec deux représentants des trois disciplines abordées : la sociologie qualitative la psychologie ergonomique et la linguistique interactionnelle. Nous complétons ces entretiens avec une brève revue de la littérature traitant de cette question.

Comme la linguistique de corpus s'est illustrée comme la discipline de référence pour la gestion des corpus nous introduirons cet article par l'examen de la place du corpus dans cette discipline avant de développer l'apport des trois autres disciplines.

1. La linguistique de corpus

Les études sur corpus en linguistique se caractérisent généralement par une approche quantitative, sur de grandes masses de données, avec des méthodes (semi)-automatiques qui visent à assurer la reproductibilité, la validation et la généralisation des résultats. Les spécialistes s'accordent à dire qu'il existe un lien entre la linguistique de corpus et l'outil informatique qui fait partie intégrante de la démarche empirique prônée par les défenseurs de la linguistique de corpus. (Péry-Woodley M.P. 1995), (Habert B. et all. 1997).

La linguistique de corpus insiste sur le caractère restrictif du corpus : « Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques *et extra- linguistiques* explicites pour servir d'échantillon *d'emplois déterminés d'une langue* » (Habert B. 2000).

En fonction des besoins de recherche, plusieurs distinctions parcourent la discipline : corpus de textes *versus* d'échantillons, corpus de référence (qui vise à représenter toutes les variétés pertinentes d'une langue) *versus* corpus de spécialité (qui est restreint à une situation de communication, un domaine, une langue de spécialité).

La constitution d'un corpus de faits langagiers soulève nécessairement le problème de la représentativité. On s'accorde aujourd'hui encore sur la difficulté, en matière de langages, à donner une « définition positive de la représentativité » (Habert B. et all 1997). Les exigences à respecter dans tous les cas sont formulées ainsi : « la taille des données doit être suffisante (par

souci de représentativité), elles doivent être diversifiées et leur origine doit être clairement mémorisée » (Habert B et all. 1998).
Pour les linguistiques de corpus, il est nécessaire de créer des ressources linguistiques communes et réutilisables. En conséquence, une attention particulière est portée à la documentation de chaque ressource. En amont, les chercheurs s'accordent sur certaines décisions relatives aux variations de domaines et de registres, sur les types de textes, sur le balisage de la structure du texte et sur l'analyse envisageable.
Une fois le corpus constitué, il fera objet d'une série de traitements le plus souvent automatiques de normalisation, de nettoyage, de segmentation, de regroupement, d'étiquetage et d'annotation (Habert B. et all. 1997, 1998). Le plus souvent sont utilisés les corpus étiquetés (chaque mot est assorti d'une étiquette) et les corpus arborés (munis d'arbres syntaxiques).
Les objectifs principaux poursuivis par la linguistique de corpus sont la constitution de ressources, leur mise à la disposition et leur confrontation par la communauté des linguistes. (Péry-Woodley M.P. 2005).

2. La sociologie qualitative

En sociologie, la notion de corpus renvoie à la méthode dite qualitative. Cette contribution n'évoque par conséquent pas la sociologie quantitative (Fox W. 1999). Typiquement, un corpus se compose d'une série de transcriptions d'entretiens, donc d'interactions occasionnées en vue de répondre à une question de recherche (Bourdieu P. 1994). Les analyses sociologiques portent également sur d'autres sources empiriques que les entretiens :

- des notes de terrains ainsi que l'expérience du sociologue ayant procédé à une observation participante.
- des documents écrits, comme, entre autres, les règlements de travail (Foucault M. 1975), les manuels de management (Boltanski L. 1999), les inscriptions publiques (Heinich N. 1995), les lettres (Boltanski L. 1984), les coupures de presse (Chateauraynaud F. 1999, Duret P. 2001), les articles scientifiques (Latour B. 1995), les affiches (Latour 1992), les manuels de savoir vivre (Elias N. 2003), les modes d'emploi (Akrich M. 1991), les listes de courses (Conein B. 1994) ou les programmes politiques,
- des photographies (Bourdieu P. 1979, Trepos 1997, Latour 1985].
- des vidéogrammes (Lee J. 1993).

La constitution du corpus de référence à partir du terrain soulève la question de sa représentativité. La réponse sociologique à cette question procède par l'épuisement de la diversité du matériau recueilli. Comme en linguistique de corpus, représentativité rime avec exhaustivité. Mais, contrairement à la fixation statistique d'un échantillon en linguistique de corpus, la logique de saturation des différences n'a aucun rapport géométrique avec la population globale (le corpus potentiel). En tant que recensement de toutes les singularités, elle est encore distincte du travail du linguiste interactionnel sur des exemplaires typiques choisis pour leur centralité.
La position du chercheur lors de ce recueil va de la neutralité distanciée de l'entretien (similaire à la posture témoignée en psychologie clinique) à l'immersion par observation participante dans un milieu écologique (comparable à celle des conversationnalistes).
L'exploitation du corpus de référence procède par relevé et analyse du déroulement et du contenu des interactions, peu (ou pas) d'attention est accordée à leur forme d'expression (contrairement aux préoccupations des linguistes). La langue (qu'elle soit orale ou écrite) n'est abordée qu'en tant que médiatrice – plus ou moins fidèle – des idées, des opinions ou des attitudes. La sociologie s'intéresse le plus souvent à « ce que raconte » le corpus, qu'elle tente d'articuler à ses questions de recherche et aux cadres théoriques mobilisés. Bien qu'il s'agisse d'une démarche qui se qualifie de qualitative, les comptages sont loin d'être absents des analyses de ce type, que la méthode passe par l'analyse de contenu, l'analyse structurale des récits ou mobilise des logiciels[1] pour exploiter les documents. La récurrence d'une idée lui attribue une première importance par rapport aux hapax. Ce présupposé n'évacue bien entendu pas les éléments marginaux qui sont également mobilisés dans les analyses.

[1] Ces logiciels (relativement nombreux) regroupent entre autres les CAQDAS (NVivo, Nud*Ist, Kwalitan, The Ethnograph), Alceste, Tropes, Prospéro, Candide, Leximappe. Pour une recension, http://www.smess.egss.ulg.ac.be/lejeune/logiciels/.

3. La psychologie ergonomique

Bien qu'elle n'emploie pas le terme corpus, la psychologie ergonomique collecte des inscriptions des activités spécifiques de résolution de problème qu'elle étudie. Ces enregistrements regroupent le film de la situation observée (quand il est possible de le réaliser) et des paroles des sujets. Dans le cas des situations de résolution collective de problème, les échanges verbaux constituent simultanément un mode d'action et une trace des activités cognitives des sujets. C'est la raison pour laquelle ces situations sont particulièrement recherchées. Lorsqu'elles ne sont pas disponibles, les sujets sont invités à commenter à voix haute ce qu'ils sont en train de faire – verbalisation simultanée – ou ce qu'ils ont fait – verbalisation *a posteriori*).

La question de la représentativité de cet ensemble de matériaux trouve une réponse différente des autres perspectives. Il n'est question ni d'exhaustivité, ni d'échantillonnage, ni de saturation. C'est le statut d'expert du sujet qui rend pertinent son témoignage. A la différence de la typicité de la linguistique interactionnelle, c'est l'efficacité du comportement qui importe.

A la différence de bien d'autres courants en psychologie, la psychologie ergonomique adopte une posture clinique qui n'exclut pas une démarche d'intervention. En effet, les enquêtes du psychologue ergonome visent à enregistrer les pratiques des sujets afin de mettre en évidence ses méthodes, d'identifier des invariants et souvent de rendre le dispositif plus opérant, plus efficace. Praxéologue, il cherche à connaître pour (mieux) agir.

Pour lui, les dits des sujets sont des intermédiaires qui donnent accès aux opérations cognitives comme le raisonnement ou l'inférence (inobservables par ailleurs). La forme des dits n'est donc pas déterminante en elle-même. Une attention accrue est par contre apportée à l'articulation de ce qui est dit et fait par le sujet. La résolution d'un problème sert de dispositif de collecte adéquat par rapport à cette posture.

Lors de l'analyse des matériaux accumulés, le psychologue cherche à identifier les processus inférentiels, les raisonnements. Il emprunte les techniques de l'analyse de contenu. En complément, il mobilise également des statistiques descriptives. L'identité disciplinaire n'est sans doute pas étrangère à ce choix, la psychologie, s'étant historiquement engagée très tôt dans le recours à ces techniques. Le psychologue cherche ensuite à dégager les éventuels invariants et à modéliser les procédures (par définition, spécifiques) mises en œuvre par les sujets.

4. La linguistique interactionnelle

Depuis quelques décennies et sous l'influence des courants interactionnistes anglo-saxons « s'est affirmée de plus en plus fortement en linguistique interactionnelle l'exigence de travailler sur des corpus de données attestées comme alternative à des démarches fondées sur l'introspection de jugements des locuteurs ou sur l'élicitation de jugements des locuteurs ». (Mondada L. 2005). Cette préférence des données naturelles sur des données fabriquées par introspection, par simulation ou par expérimentation s'inscrit dans un mouvement général dans plusieurs disciplines de la nouvelle communication (Winkin Y. 1981).

Une deuxième exigence invite à travailler sur des *enregistrements* –audio ou vidéo- d'interactions sociales, c'est-à-dire sur des données permettant de documenter l'émergence et le déploiement de ces pratiques *dans le temps*. On ne travaille donc *ni* sur les descriptions de ces pratiques (dans les entretiens, dans des notes prises par le chercheur) *ni* sur des produits de celles-ci (par exemple dans des textes issus d'une manière ou d'une autre de l'activité). (Mondada L. 2005).

Le problème de la représentativité n'est pas nécessairement articulé à une ambition de généralisation. La réponse du chercheur à cette question consiste seulement à dégager ce qui est propre au corpus étudié, ce qui en fait le style ou ce qui s'y manifeste comme phénomènes récurrents (Condamines A. 2005). Le corpus devra être pertinent (Vincent D. 2003) ou de « qualité » (Plantin C. 2005) ; pour ce dernier, le corpus doit manifester les trois dimensions suivantes :

- technique : il faut réaliser des corpus de bonne qualité technique sonore et visuelle pour faciliter leur conservation, leur transcription et leur exploitation manuelle ou informatique
- juridique : l'enjeu juridique touche trois dimensions : le respect de la vie privée des personnes enregistrées (accord préalable des enquêtés, anonymisation des données), le droit d'auteur (entre collecteurs, transcripteurs et chercheurs) et le recueil et la diffusion de données (préparation et mise en place de l'enregistrement).
- sociolinguistique : un « bon corpus » est souvent décrit dans la littérature comme « naturel », « authentique » et « représentatif » (dans le sens évoqué ci-dessus).

Au niveau de l'analyse, le chercheur s'attelle particulièrement à décrire l'organisation, le fonctionnement et les enjeux des interactions. Il rend compte de manière parallèle des phénomènes verbaux, vocaux et gestuels. Une attention est portée à la dimension comportementale ou pragmatique des échanges dont l'unité minimale est l'acte de langage.
S'appuyant sur ces considérations méthodologiques fortes, quelques équipes internationales (comme le projet CLAPI mené par le laboratoire ICAR à Lyon) se sont engagées dans l'archivage de corpus parlés en interaction. Cet archivage a plusieurs visées :

- patrimoniale : sauvegarder des façons de parler et constituer une documentation historique des usages de la langue en interaction
- scientifique : permettre les études empiriques de phénomènes linguistiques
- appliquée : tirer des préconisations ou applications des études.

Discussion

Au terme de cette confrontation disciplinaire, nous pouvons avancer quelques constatations[1]. La posture du chercheur par rapport à son objet varie également d'une discipline à l'autre. Là où la linguistique de corpus vise une objectivation des phénomènes dont elle rend compte, le sociologue tente d'appréhender la subjectivité des différents acteurs. Le recours aux entretiens le positionne dans une neutralité distanciée par rapport à ses informateurs. Il arrive que certains sociologues procèdent à une observation participante. Ils tentent alors une immersion dans le phénomène social qu'ils observent. En linguistique interactionnelle, il existe une double tendance. Certaines recherches impliquent la conversion du chercheur qui acquiert, par immersion ou imprégnation, la compétence des locuteurs étudiés (ce qui est proche de l'observation participante en sociologie ou en anthropologie). D'autres recherches procèdent à l'effacement du dispositif de recueil des données. Dans ce deuxième courant, le chercheur est alors un pur observateur, distant et, idéalement, invisible. Cette position est similaire à la posture clinique témoignée par le psychologue ergonomique. Ce dernier ne s'imprègne pas du terrain (vu qu'il n'apprend pas à effectuer les tâches) mais, vu qu'il n'exclut pas l'intervention, il n'emprunte pas non plus la neutralité distanciée des sociologues.
En bref, la sociologie qualitative procède typiquement par entretiens. La psychologie ergonomique s'intéresse aux situations de résolution de problème. La linguistique interactionnelle privilégie une démarche naturelle d'observation, portant sur des situations qui ne sont pas créées, arrangées, préparées par le chercher pour les fins de son enquête. Le comportement interactionnel est enregistré dans sa totalité verbale et non verbale.
Ces disciplines traitent différemment des dires, des actes et de leur contexte. A la différence de la perspective linguistique, le psychologue et le sociologue s'intéressent plus au contenu des dires qu'à leur forme. Toutefois, là où le sociologue s'intéresse surtout à ce que l'informateur lui dit, le psychologue ergonomique étudie l'articulation du geste et de la parole. Le conversationnaliste prolonge plus loin encore cette attention.
La façon dont les interactions sont transcrites est congruente avec ces différentes démarches : le sociologue transcrit la teneur du propos (ce qui est dit), sans trop se préoccuper (la plupart du temps) des gestes, postures, intonations, pauses et hésitations. Le psychologue ergonomique sera lui particulièrement attentif à transcrire en parallèle le geste et le discours. A travers l'analyse conversationnelle, la linguistique interactionnelle s'est enfin particulièrement illustrée pour ses transcriptions fines et détaillées incorporant les chevauchements, les pauses (chronométrées), les intonations et le non-verbal.
Plus préoccupée de la signification du discours que de sa forme, la psychologie et la sociologie partagent plusieurs de leurs outils d'exploitation et d'analyse du corpus (avec, en bonne position, l'analyse de contenu). A nouveau, cependant, des différences se manifestent dans l'usage de ces techniques. Le sociologue se contente de recenser les différents arguments, idées, logiques ou phases d'un entretien. L'épuisement de la diversité lui suffit. Le psychologue sera, pour sa part, plus enclin à recourir à des techniques d'objectivation comme les statistiques. En linguistique interactionnelle, l'analyse ne se limite pas au contenu, elle prend pour sa part également en considération le niveau formel, à travers la structure et l'organisation des interactions. L'unité d'analyse sera, en fonction du niveau de granularité, la conversation, la séquence, l'échange, l'intervention ou l'acte de langage.

[1] Une présentation plus détaillée sera proposée lors de la communication orale.

En ce qui concerne la question de la représentativité, chaque discipline propose une réponse singulière. La sociologie fait rimer représentativité avec exhaustivité. Si cette acception la rapproche de la linguistique de corpus, la logique de saturation qu'elle emprunte la distingue cependant de la fixation statistique d'un échantillon. L'épuisement des diversités n'a en effet aucun rapport géométrique avec la population globale. La psychologie ergonomique et la linguistique interactionnelle donnent à cette même question de la représentativité une réponse qui ne s'articule pas toujours à une visée d'exhaustivité. Les raisons de chacune de ces disciplines sont cependant différentes. Vu que la psychologie ergonomique s'attelle à enregistrer les procédures efficientes, c'est le statut d'expert du sujet qui l'emporte sur le nombre de sujets rencontrés. La linguistique interactionnelle travaille, pour sa part, sur des exemplaires typiques extraits des phénomènes récurrents et réguliers observés dans les interactions.
En conclusion, cette confrontation nous a permis de prendre conscience des écarts disciplinaires. Aucune discipline ne peut à elle seule épuiser la complexité des phénomènes interactionnels. On s'interrogera sur les conditions de (1) partage de corpus commun entre experts de ces différentes disciplines, (2) leur analyse collective et (3) d'éventuelles transformations des pratiques de chacun pour rendre cette collaboration transdisciplinaire opérante.

BIBLIOGRAPHIE

AKRICH, M. et BOULLIER, D. 1991. Le mode d'emploi : genèse, forme et usage, in D. Chevalier (éd.), *Savoir faire et pouvoir transmettre. Transmission et apprentissage des savoir-faire et des techniques*, Maison des sciences de l'homme.
BILGER, M. (éd). 2000. *Corpus : Méthodologie et applications linguistiques*, Paris, Honoré Champion.
BOLTANSKI, L. et CHIAPELLO E. 1999. *Le nouvel esprit du capitalisme*, Paris, Gallimard.
BOLTANSKI, L., DARRÉ, Y., et SCHILTZ M.-A. 1984. La dénonciation, *Actes de la Recherche en sciences sociales*, 51, pp. 3-40.
BOURDIEU, P. 1979. *La distinction. Critique sociale du jugement*, Paris, Minuit.
BOURDIEU, P. 1994. *Raisons pratiques. Sur la théorie de l'action*, Paris, Seuil.
CONDAMINES, A. (dir.) 2005. *Sémantique et corpus*, Londres, Hermès.
CHATEAURAYNAUD, F. et TORNY, D. 1999. *Les sombres précurseurs. Une sociologie pragmatique de l'alerte et du risque*, Paris, École des Hautes Études en Sciences Sociales.
CONEIN, B. et JACOPIN, E. 1994. Action située et cognition. Le savoir en place, *Sociologie du Travail*, 4, pp. 475-500.
DURET, P. et TRABAL, P. 2001. *Le sport et ses affaires. Une sociologie de la justice de l'épreuve sportive*, Paris, Métailié.
ELIAS, N. 2003. *La Civilisation des moeurs*, Paris, Agora.
ERICSSON, K.A., SIMON, H.A. 1993. *Protocol analysis : Verbal reports as data*. MIT, Cambridge.
FOUCAULT, M. 1975. *Surveiller et punir. Naissance de la prison*, Paris, Gallimard.
FOX, W. 1999. *Statistiques sociales*, Paris, De Boeck.
HABERT, B. 2000. Des corpus représentatifs : de quoi, pour quoi, comment ?, in M. Bilger (éd.), *Linguistique sur corpus. Études et réflexions*, Perpignan, Presses Universitaires de Perpignan, pp.11-58.
HABERT, B., NAZARENKO, A., SALEM, A. 1997. *Les linguistiques de corpus*, Paris, Armand Colin/Masson.
HABERT, B., FABRE, C., ISAAC, F. 1998. *De l'écrit au numérique : constituer, normaliser et exploiter les corpus électroniques*, Paris, InterEditions.
HEINICH, N. 1995. Les colonnes de buren au palais-royal. ethnologie d'une affaire, *Ethnologie française*, XXV, 4, pp. 525-541.
HOC, J.-M., DARSE, F. (éds.) 2004. *Psychologie ergonomique : tendances actuelles*, Paris, PUF.
KERBRAT-ORECCHIONI, C. 1994. *Les interactions verbales*, Paris, Armand Colin.
LATOUR, B. 1985. Les "vues" de l'esprit. Une introduction à l'anthropologie des sciences et des techniques, *Culture Technique*, 14, pp. 5-29.
LATOUR, B. 1992. *Aramis ou l'amour des techniques*, Paris, La Découverte.
LATOUR, B. 1995. *La Science en action*, Paris, Gallimard [La Découverte, 1989].
LEE, J. et WATSON, R. 1993. Regards et habitudes des passants. Les arrangements de visibilité de la locomotion, *Les annales de la recherche urbaine*, 57-58, pp. 101-109.

MONDADA, L. 2005. L'analyse de corpus en linguistique interactionnelle : de l'étude de cas singuliers à l'étude de collections, in A. Condamines (dir.),*Sémantique et corpus*, Londres, Hermès.
PERY-WOODLEY, M.-P. 1995. Quels corpus pour quels traitements automatiques ?, TAL, 36(1-2), pp. 213-232.
PERY-WOODLEY, M.-P. 2005. Discours, corpus, traitement automatiques, in A. Condamines (dir.), *Sémantique et corpus*, Londres, Hermès.
PLANTIN, C. 2005. Pour une archive des langues parlées en interaction. Statuts juridiques, formats et standards, représentativité, in J.L. Lebrave, *La société de l'information et ses enjeux, actes du colloque de bilan « la société de l'information »* 2001-2005, ENS-LSH, Lyon.
PLETY, R. 1993. *Ethologie des communications humaines. Aide-mémoire méthodologique*, Lyon, Arci-PUL.
RASTIER, F. (éd) 2000. *Une introduction aux sciences de la culture*, Paris, PUF.
RASTIER, F. 2005. Les enjeux épistémologiques de la linguistique de corpus, in G. Williams (éd.), *La linguistique de corpus*, Rennes, Presses universitaires de Rennes.
RICHARD, J.-F. 2005. *Les activités mentales. Comprendre, raisonner, trouver des solutions*, Paris, Armand Colin.
TRAVERSO, V. 1999. *L'analyse des conversations*, Paris, Nathan (coll. 128).
TREPOS, J.-Y. 1997. Approche méthodologique de l'utilisation de la photographie dans l'enquête sociologique, Notes du séminaire du 17 novembre 1997.
WINKIN, Y. 1981. *La Nouvelle Communication*, Paris, Éditions du Seuil.

PORPHYRY AU PAYS DES PAESTANS
Usages d'un outil d'analyse qualitative de documents par des étudiantes de maîtrise en iconographie grecque

Aurélien BÉNEL
Laboratoire Tech-CICO (Institut Charles Delaunay)
Université de Technologie de Troyes

SOMMAIRE

Résumé : *Cette communication livre un retour d'expérience sur l'introduction de Porphyry, logiciel d'annotation sociale, dans l'activité d'analyse d'un corpus de photographies par une équipe d'historiens d'art. Leur corpus comprend les photographies (profil, détail du recto, détail du verso) des 300 vases de la région de Paestum (en Campanie) représentant des scènes de banquet ou liées à Dionysos. Au printemps 2005, trois étudiantes ont contribué à l'étude iconographique de ces vases, dans le cadre de leur maîtrise en sciences de l'antiquité.*
Le compte-rendu de cette expérience par le responsable scientifique de l'équipe nous intéresse à plus d'un titre : par la description de l'activité instrumentée d'annotation dans un « jargon » différent de celui des concepteurs du logiciel, par l'absence de mention à des fonctionnalités jusqu'alors considérées comme capitales, enfin, par des critiques, en apparence anodines, qui pourraient remettre en question certaines positions théoriques.
Par exemple, pourquoi regretter le fait que l'on ne puisse pas « renommer un descripteur » ? Sans être pour autant un concept, serait-il tout de même plus qu'une chaîne de caractère ? Le mot « descripteur » est-il alors bien choisi ? De même, pourquoi dénoncer l'absence de comptage automatique dans un outil qui se veut avant tout qualitatif ? D'ailleurs, que s'agit-t-il de compter : les fragments, les sources photographiques, ou bien, indirectement, les vases ? N'y a-t-il pas alors un surprenant « glissement » référentiel ? Tel est le genre de questions qui seront abordées dans cette communication.

Remerciements : *Cet article constitue le premier retour d'expérience d'une collaboration de trois ans entre le réseau ARTCADHi-CNRS[1] et le laboratoire CRATA[2]. Cette collaboration n'a pu être possible que grâce au travail et à la persévérance des animateurs scientifiques et de leurs équipes. Nous remercions particulièrement Andrea Iacovella, Jean-Marc Luce, Véronique Pouyadou, Samuel Gesche, Pascale Jacquet et ses étudiantes.*

1. Introduction
Comme aurait pu le dire *Qohéleth*, il y a un moment pour tout et un temps pour chaque chose dans la recherche en informatique : un temps pour lire et un temps pour écrire, un temps pour théoriser et un temps pour développer des logiciels, un temps pour former les usagers et un temps pour les observer travailler, un temps pour mettre à l'épreuve les hypothèses et un temps pour en forger de nouvelles.
Dans cette communication, nous allons donner un retour d'expérience sur l'introduction de *Porphyry*, logiciel d'annotation sociale [Iacovella *et al.*, 2005 ; Bénel, 2003], dans l'activité d'analyse d'un corpus de photographies par une équipe d'historiens d'art. Le corpus comprend les photographies (profil, détail du recto, détail du verso) des 300 vases de la région de Paestum (en Campanie) représentant des scènes de banquet ou liées à Dionysos [Pouyadou, 2001]. Au

[1] Réseau ARTCADHi-CNRS : « Atelier de recherches transdisciplinaires sur la construction du sens en archéologie et autres disciplines historiques », <http://www.porphyry.org/>.
[2] Laboratoire CRATA : « Culture, représentations, archéologie, théâtre antique », <http://www.univ-tlse2.fr/crata/>.

printemps 2005, trois étudiantes ont contribué à l'étude iconographique de ces vases, dans le cadre de leur maîtrise en sciences de l'antiquité.
Le compte-rendu de cette expérience par le responsable scientifique de l'équipe nous intéresse à plus d'un titre. Par exemple, l'usage du nom des primitives de notre outil (*descripteurs* : *facettes*, *descripteurs ordinaires*, *identifiants*), usage souvent difficile, parfois « fautif », nous amènera à nous questionner sur notre « jargon » et à le simplifier. De même, l'absence de mention à des fonctionnalités jusqu'alors considérées comme capitales pourra nous amener à nous interroger sur ce qui les rend inopérantes. L'analyse de critiques, en apparence purement techniques, pourra également remettre en question certaines positions théoriques.
Pour analyser ce compte-rendu d'usage, nous prendrons pour grille les quatre objets principaux autour desquels l'activité instrumentée se déploie : le corpus, la source, le point de vue et le portfolio.

2. Corpus

Les photographies de chaque vase ayant été numérisées, l'expert les importe dans le système pour qu'elles deviennent des *sources.* Survient alors un premier problème signalé dans le compte-rendu :

> Inconvénients : […] 2 - L'impossibilité d'introduire dans le nom d'un descripteur un mot qui figure dans le nom de la facette. (Luce, 2005)

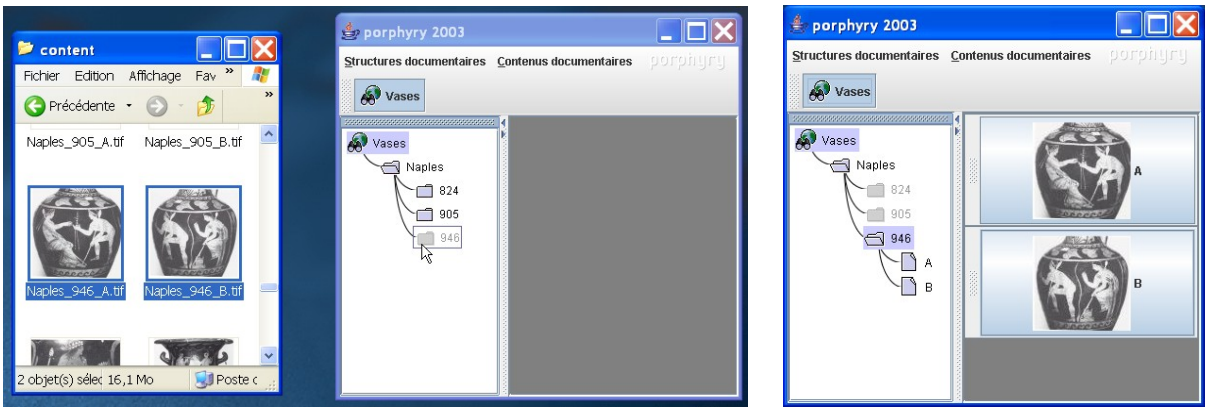


Fig. 1 : Séquence d'importation de sources (par un glisser-déposer) lors de la constitution du corpus.

Au moment de l'import, un *identifiant d'objet documentaire* est assigné à la source. C'est cet identifiant qui permettra à différents experts de savoir qu'ils parlent de la même source. Par conséquent, l'identifiant doit correspondre à une unique source par serveur et être pérenne.
Gênés par la valeur un peu trop objective de ces identifiants, nous avons d'abord choisi de leur donner des valeurs aléatoires[1]. Par la suite, nous avons dû faire face aux imports successifs des mêmes sources (dont les différentes analyses ne pouvaient plus être corrélées) et surtout à l'incapacité d'extraire des rapports qui soient compréhensibles à l'extérieur du système.
L'identifiant aléatoire est alors remplacé par le nom du fichier importé, souvent significatif, faisant consensus et pérenne (numéro d'archive, référence bibliographique succincte…). Dans notre cas, comme d'en d'autres, la nomenclature utilisée reprend une logique de rangement hiérarchique : « Louvre_K217_B » pour la photographie B du vase K217 du Louvre. Cette nomenclature ayant une signification, il devient alors intéressant de la retrouver en tant que structure d'analyse au même titre que les points de vue des experts. La source « Louvre_K217_B » peut alors être décrite par le *réseau de descripteurs* « Louvre > K217 > B ». La création de ce réseau, tâche répétitive et sans grand intérêt pour l'expert, est censée être rendue « transparente » grâce à un algorithme simple sur les chaînes de caractères des noms de fichiers à importer, à en croire la doléance mentionnée plus haut, elle n'en est que plus obscure…
Si nous avons choisi de discuter de ce problème, en apparence trivial, c'est qu'il révèle selon nous une question fondamentale. Nous nous plaisons à considérer le texte (ou ici l'image) comme le lieu de confrontation des différentes analyses des experts. Cet objet, parce qu'il est toujours à

[1] L'identifiant d'objet documentaire reste d'ailleurs, aujourd'hui encore, caché de l'usager.

interpréter, nous éviterait les pièges de tous les intégrismes (le positivisme y compris), et, parce qu'il est accessible à tous et qu'il devient « document » (au sens de « preuve »), nous protègerait du relativisme. Le nom donné au document, toujours relatif à d'autres noms et donc à la structure du corpus, devient alors extrêmement problématique. En effet, si le document joue bien un rôle de « butée ontologique » vis-à-vis des points de vue d'experts, comment ne pas retrouver dans la structure du corpus la naïveté des « ontologies informatiques », supposées référentielles et éternelles ?

3. Source

Une fois le corpus établi, même partiellement, le travail sur les sources peut commencer. Le responsable souligne :

> Avantages : […] 8 - Système moins contraignant que le travail sur des tableaux. Le système maintient l'opérateur en constant contact avec les images. Les étudiantes ont souligné le caractère ludique des manipulations. (Luce, 2005)

Parmi les manipulations permises par le système sur les sources, une seule laisse des traces : celle consistant à sélectionner des fragments. Or, il n'en est fait aucune mention dans le compte-rendu. L'observation du corpus nous révèle d'ailleurs une absence totale de fragments. Comment expliquer que les usagers d'un logiciel d'annotation n'utilisent pas une fonctionnalité *a priori* aussi centrale que la gestion des fragments ? Formation à l'outil trop courte ? Ergonomie imparfaite de la création et de l'utilisation des fragments ? Des éléments de réponse nous sont donnés lorsque l'on cherche un exemple dans le corpus de ce que l'emploi de fragments aurait pu apporter.

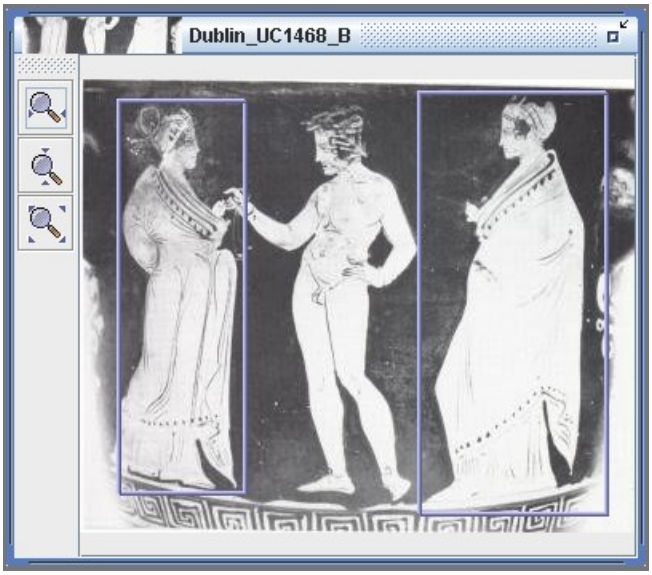


Fig. 2 : Cas fictif de sélection dans une source de fragments répondant à la même thématique.

Tout d'abord, il apparaît que les éléments graphiques intéressant nos experts, par exemple un oiseau et une chèvre, sont tout à fait visibles dans la scène d'ensemble et, dans la majorité des cas, présents une seule fois. Quel intérêt alors de les sélectionner ? Nous nous trouvons sans doute ici devant une curiosité due au matériel étudié lui-même : les scènes représentées sur les vases semblent utiliser les figures avec parcimonie comme pour accentuer leur valeur symbolique.
Ensuite, on serait en droit de s'interroger sur l'intérêt d'afficher, par exemple, la position d'un personnage à la jambe levée sans afficher *en même temps* celle du personnage assis qui lui fait face. Autrement dit, la sélection de fragments ne semble avoir d'intérêt que dans une logique différentielle.
Enfin, on peut douter du bien-fondé d'une l'interface qui encourage la création de catégories d'analyse, nommées dès le départ, puis la création de fragments dans ces catégories. Sans doute faudrait-il plutôt permettre dans un premier temps de sélectionner les fragments, de les rassembler ensuite en groupes anonymes et aux limites un peu floues, et de préciser ces limites graduellement.

4. Point de vue

Les ensembles de fragments (ou de sources) vont ensuite être décrits par l'expert, à l'aide de *descripteurs.* Cette appellation avait pour but d'éviter les « concepts » chers à l'intelligence artificielle et d'indiquer que ces simples termes avaient pour seule vocation d'être utilisés dans des descriptions. L'emprunt de ce nom aux sciences documentaires est cependant malheureux puisqu'il évoque des « listes d'autorité », consensuelles et pérennes, là où au contraire notre descripteur est propre au *point de vue* de l'expert (sorte d'idiolecte ou de sociolecte), et peut y apparaître ou disparaître à tout moment.
Mais si le descripteur est un terme, comment interpréter le commentaire suivant ?

> Inconvénients : 1 - L'impossibilité de renommer un descripteur. C'est très ennuyeux en cas d'erreur. On est obligé de refaire le descripteur. (Luce, 2005)

Plutôt que par son nom, le « descripteur » ne serait-il pas davantage caractérisé, comme le laisse entendre le commentaire suivant, par les fragments et les sources qu'il contient déjà ?

> Avantages : […] 2 - Possibilité de faire des liens avec l'apparition immédiate des images. (Luce, 2005)

Tout porte à croire que nous sommes en présence d'une *collection* (Rousseaux, 2006), qui, à l'image de la collection d'œuvres d'art, serait construite, non à partir d'une définition subsumant chacun de ses membres, mais en traçant un parcours à travers leurs différences. La constitution de cette collection pourrait alors se faire dans la continuité des regroupements dont nous parlions dans la section précédente. La collection permettrait de donner une frontière à ce groupe, puis éventuellement un nom, le tout de manière transitoire au gré des découvertes de singularités parmi les sources et les fragments.

Une fois créées, les collections peuvent être organisées entre elles :

> Avantages : 1 - La possibilité de créer des rubriques à l'infini pour réaliser un relevé iconographique précis. Cette possibilité est très appréciable par rapport aux bases de données. (Luce, 2005)

> Inconvénients : […] 5 - Quand les liens sont nombreux, l'arborescence devient vite illisible. (Luce, 2005)

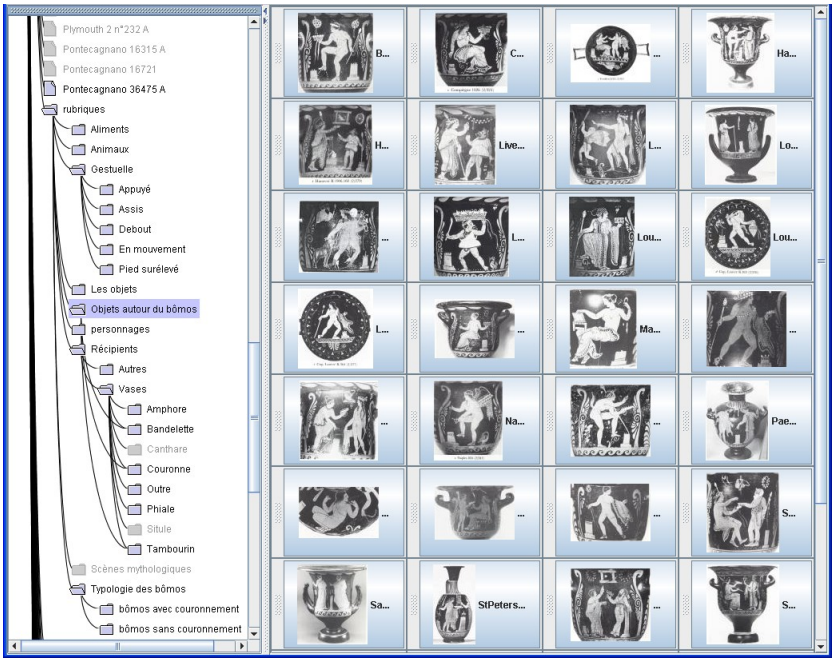


Fig. 3 : Point de vue portant sur l'étude des scènes où figurent des autels.

Le réseau des collections comporte les inconvénients et les avantages de sa complexité. Poly-hiérarchique, il permet à un même objet documentaire d'être décrit selon plusieurs dimensions

d'analyse et même à une collection d'être à l'intersection de plusieurs collections (ce qui est particulièrement utile pour le temps, l'espace, ou encore les rôles d'un objet). Une telle structure l'empêche par contre d'être affiché dans son ensemble sans que ses arcs ne se coupent. Une visualisation interactive pourrait être mise en place de sorte que seuls certains arcs soient affichés en réponse aux actions de l'usager.

5. Portfolio

Corpus et points de vue sont rassemblés dans un portfolio. C'est le moment de feuilleter ce corpus enrichi, le moment de découvrir le corpus sous un jour nouveau.

> Avantages : […] 4 - Créer des associations auxquelles on n'aurait pas pensé ; 5 - Le logiciel est particulièrement intéressant pour mettre en évidence des cas exceptionnels. (Luce, 2005)

Les commentaires que nous venons de citer font référence à la possibilité de sélectionner plusieurs collections et de voir en retour, d'une part, les objets documentaires présents à leur intersection, d'autre part, le statut de co-occurrence des autres collections (disjonction, conjonction partielle, conjonction totale).

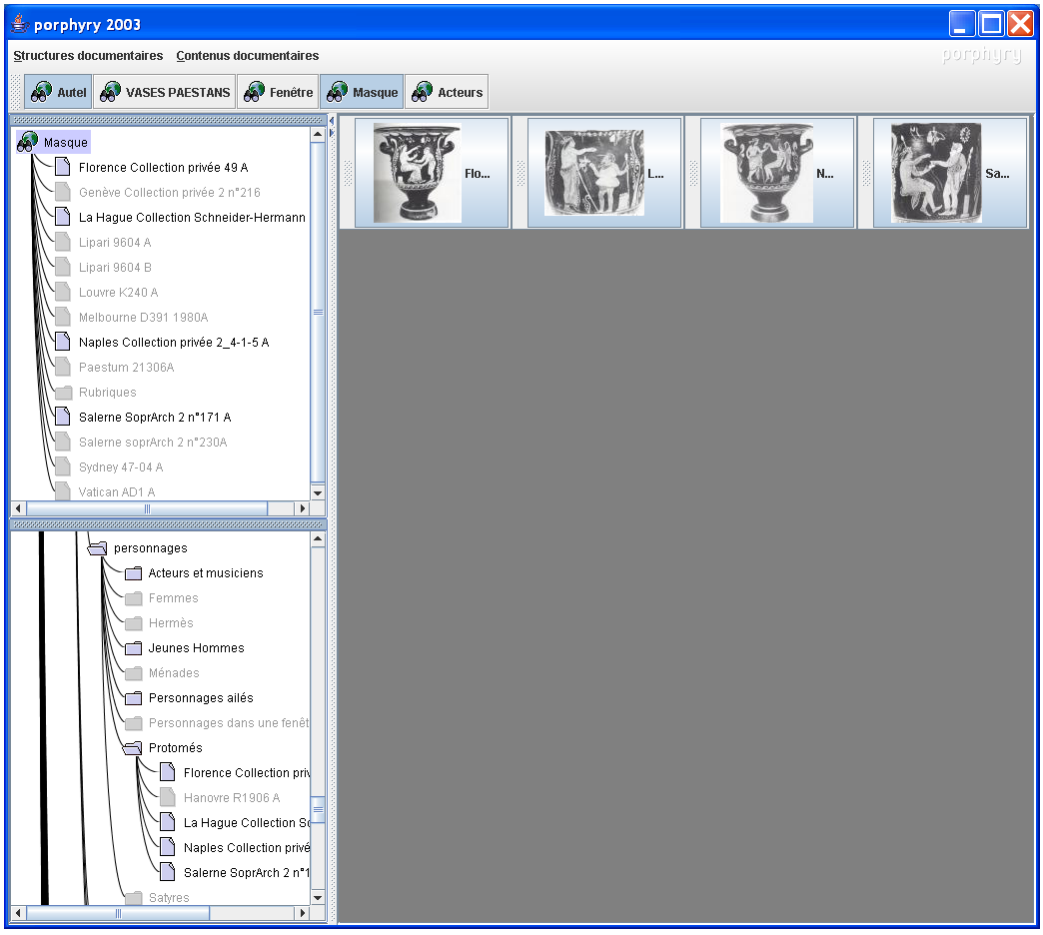


Fig. 4 : Statut de co-occurrence entre les masques et les différents types de personnage autour de l'autel.

> Inconvénients : […] 4 - On ne peut pas demander à l'opérateur de constamment compter les vignettes qui s'affichent. Il faut un outil qui les compte automatiquement. (Luce, 2005)

Pourquoi dénoncer l'absence de comptage automatique dans un outil qui se veut avant tout qualitatif ? Ne serait-ce pas par souci de l'exhaustivité ? Le même outil, dans les mains d'historiens de la Chine des années 30, est utilisé au contraire pour relever une ou deux manifestations typiques d'un phénomène (habillement, technique, etc.). Il n'est pas impossible que cette différence de statut donné aux sources dépende de la discipline. Il est intéressant de noter que l'interface homme-machine actuelle n'affiche pas l'ensemble des objets documentaires mais juste

quelques exemples prototypiques. Les usagers ont cependant réussi à contourner ce comportement du logiciel en dupliquant à la racine du point de vue chaque objet documentaire.
Par ailleurs, que s'agit-t-il de compter : les fragments, les sources photographiques, ou bien, les dossiers d'archive et donc, indirectement, les vases ? N'y a-t-il pas alors un surprenant « glissement » référentiel ? Ceci vient rappeler le rôle de *témoin* qu'assure le document pour la communauté scientifique en archéologie vis-à-vis d'objets et de structures rendus à jamais inaccessibles par l'analyse destructive qu'est la fouille (*Iacovella et al*., 2005).
Les commentaires suivants mettent l'accent sur le caractère social du portfolio, lieu de publication (au sens de « rendre public ») et de confrontation des points de vue, lieu d'accord et de désaccord, de sédimentation et de révolution dans les savoirs.

> Avantages : […] 3 - la possibilité à l'avenir de rentrer plusieurs travaux, chacun peut ensuite bénéficier du travail des autres ; […] 6 - à partir d'un même corpus de base, possibilité de créer plusieurs facettes sur des thèmes différents, 7 - possibilité de développer des réflexions différentes sur un même corpus. (Luce, 2005)

> Au fur et à mesure que de nouveaux groupes travailleront sur porphyre, les corpus vont s'élargir, les liens vont devenir plus nombreux et il sera possible de comparer des points de vue en utilisant les diverses facettes. (Luce, 2005)

> Inconvénients : […] 6 - L'impossibilité d'imprimer autrement que par la capture d'écran qui ne permet pas d'imprimer tout ce qui est sélectionné si le nombre de descripteurs ou des documents ne tient pas dans l'écran, 7 - l'impossibilité de sortir du bureau pour utiliser porphyre, même pour montrer les résultats aux collègues. (Luce, 2005)

Le dernier point évoqué pourrait surprendre. Pourquoi vouloir à tout prix imprimer les résultats et les « sortir du bureau », quand on utilise un environnement visant une continuité numérique entre la lecture et l'écriture ?
On pourrait tout d'abord objecter que « sortir du bureau pour utiliser porphyre » ne sera bientôt plus un problème, les réseaux sans fil se généralisant. Mais le problème est peut-être plus profond : la publication ne nécessiterait-elle pas un changement de medium ? Si l'on peut envisager des « passerelles » vers le Web et le monde du papier, comment le faire sans perdre les propriétés du portfolio numérique ? A moins que ce soit justement le caractère plus stable et moins polémique qui soit recherché dans ces média. Ne risque-t-on pas alors de passer à côté de la révolution d'une herméneutique numérique ?

6. Conclusion

Le parcours que nous venons de tracer à travers un compte-rendu de l'usage de *Porphyry*, nous a permis de poser un certain nombre de questions portant sur les conditions de construction du sens, à plusieurs, à partir de documents. Cette communication est l'occasion rêvée de poser ces questions à d'autres et d'obtenir en retour quelques éléments de réponse. Nous ne doutons pas que ces éléments de réponse configureront ce que seront dans le futur ce type d'outils d'analyse qualitative de documents et plus particulièrement leurs interfaces homme-machine.
La prochaine étape souhaitée serait l'appropriation de l'outil dans un usage quotidien.

> Elles [les étudiantes] ont conclu en exprimant leur frustration sur un point : leur mémoire de maîtrise a été fait à l'aide de tableaux. Maintenant qu'elles ont goûté à Porphyre, elles souhaitent pouvoir s'en servir pour leurs propres mémoires. Elles le feront pour leur mémoire de master 2. (Luce, 2005)

Espérons que l'appellation de l'outil par un nom (« Porphyre » et même « porphyre ») différent de son nom officiel (« Porphyry ») soit le gage d'une appropriation déjà en marche !

BIBLIOGRAPHIE

BÉNEL, A. 2003. Consultation assistée par ordinateur de la documentation en Sciences Humaines : Considérations épistémologiques, solutions opératoires et applications à l'archéologie, Thèse de doctorat en informatique, INSA de Lyon, décembre 2003. In : *Texto ! mars 2004*. Disponible sur : <http://www.revue-texto.net/Inedits/Benel/Benel.html>.

IACOVELLA, A., BÉNEL, A., CALABRETTO, S., HELLY, B. 2005. Assistance à l'interprétation dans les bibliothèques numériques pour les sciences historiques, in J.-L. Lebrave (éd.), *La société de l'information et ses enjeux, Actes du colloque de bilan du programme interdisciplinaire « Société de l'information »,* 2005. pp. 167-179.
Disponible sur : http://www.porphyry.org/Members/abenel/benel_PSI_05.pdf
LORTAL, G., LEWKOWICZ, M., TODIRACU-COURTIER, A. 2005. Modélisation de l'activité d'annotation discursive pour la conception d'un collecticiel support à l'herméneutique, in *Actes des 16ème journées francophones d'ingénierie des connaissances*, Grenoble, PUG, pp.169-180.
LUCE, J.-M. 2005. *Porphyre à Toulouse*, courriel adressé à A. Iacovella et A. Bénel, 29 juin 2005.
POUYADOU, V. 2001. Dionysos barbu : le sens du poil, *Pallas 57*, pp. 169-183.
ROUSSEAUX, F. 2006. La collection, un lieu privilégié pour penser ensemble singularité et synthèse, *EspacesTemps.net, février 2006*.
Disponible sur :<http://espacestemps.net/document1836.html>

LA RÉSURRECTION DU DICTIONNAIRE ANCIEN PAR LA DÉCONSTRUCTION POSITIVE DE L'INFORMATIQUE

Christophe REY & Corinne ZAOUI
Université de Provence, Equipe DELIC

SOMMAIRE

Résumé : *Les travaux d'étude et d'informatisation des dictionnaires anciens que poursuit l'un d'entre nous (C. REY) mettent clairement en évidence les problèmes d'informatisation liés à la nature même de ces données, à savoir un certain manque de rigidité structurelle. Ce manque de rigidité occasionne indubitablement une moins bonne lisibilité des données, que les dictionnaires modernes – par le biais de leur informatisation systématique – semblent avoir réglée. Il est ainsi parfois difficile d'identifier les différents champs d'information constitutifs des tout premiers dictionnaires, et de repérer précisément, par exemple, ce qui relève du marquage grammatical, de la définition, etc., afin de pouvoir reconstruire, récupérer, retrouver ou mettre en évidence les informations les plus saillantes qui s'y trouvent.*

Les soucis de structuration des données pour une manipulation et une interrogation grâce aux normes de codage informatique nouvelles prônées par C. ZAOUI, permettent d'apporter des solutions nouvelles à notre projet de rétroconversion d'un dictionnaire encyclopédique, le dictionnaire Grammaire & Littérature (1782-1786) de l'Encyclopédie Méthodique (1782-1832).

L'un de nos objectifs communs étant de proposer une solution qui, à travers ses choix méthodologiques et informatiques, soit généralisable, ou tout au moins facilement réexploitable sur d'autres corpus, nous avons donc fait émerger le concept d'un balisage XML "souple" ou "flottant".

La solution des balisages précédents, soit minimal (WOOLDRIDGE 1994 et 1996, LEROYTURCAN, 1996) soit analytique (WIONET et TUTIN) n'étant pas pleinement satisfaisante, cette notion de balisage "souple" nous permettait d'allier la souplesse du premier et la richesse du second, tout en respectant la nature non rigide du document et en autorisant ainsi un balisage identifiant les grands champs informationnels et en mettant en évidence un certain nombre d'informations flottant à l'intérieur de ces grands constituants. L'efficacité de notre type de balisage a pu être mise à l'épreuve à travers la réalisation d'un outil d'interrogation de fichiers XML : CorpXML (http://www.up.univmrs.fr/delic/perso/rey/methodique/index.htm).

Introduction

Ainsi que l'illustre le foisonnement des ouvrages "grand public" paraissant à la fois sous leur forme papier et sous une forme informatique, le dictionnaire électronique semble s'être installé assez largement dans le panorama lexicographique actuel. Pour s'imposer, ce dernier a bénéficié non seulement de la place grandissante de l'outil informatique dans les techniques du monde de l'édition, mais aussi de l'aspect très structuré de nos dictionnaires modernes. L'évocation de cette dynamique très active d'informatisation des dictionnaires ne peut légitimement être faite sans que soit mise en avant la question très particulière de l'informatisation des dictionnaires anciens.

Initiés depuis les années 80 par le lexicographe Russon Wooldridge, les travaux de mise à disposition électronique des monuments lexicographiques de notre passé se distinguent comme une entreprise répondant à des contraintes bien particulières. La communication que nous proposons s'inscrit dans la lignée des études déjà existantes sur ce sujet et décrit la solution du balisage *souple* ou *flottant* pour l'informatisation des dictionnaires anciens de nature encyclopédique.

Après avoir dressé un bref rappel des solutions de rétroconversion déjà existantes, nous présentons les spécificités du balisage que nous avons mis au point et revenons sur l'application que nous en avons faite sur l'un des trente-neuf dictionnaires de matière de l'*Encyclopédie Méthodique* (1782-1832).

1. Le point sur l'informatisation des dictionnaires anciens

Répondant à des caractéristiques macrostructurelles et microstructurelles qui ne caractérisent pas les dictionnaires modernes, les dictionnaires anciens constituent, en vue de leur informatisation, un objet d'étude particulièrement délicat. Parmi les nombreux projets de recherche consacrés à cette question de la rétroconversion des dictionnaires des siècles précédents, deux grandes solutions d'informatisation semblent s'être imposées : celle du balisage *minimal* et celle du balisage *analytique*.

1.1. Les méthodologies de rétroconversion

Les deux grandes solutions que constituent le balisage *minimal* et le balisage *analytique* possèdent chacune leurs spécificités.

Utilisée pour la rétroconversion du *Dictionnaire de l'Académie Française* (1694) (Wooldridge, 1994), ou du *Dictionnaire Critique* (1787) de Jean-François Féraud (Caron, Dagenais, Gonfroy, 1992), la solution du balisage *minimal* est une approche minimaliste guidée par le souci de préserver l'intégralité du texte informatisé. Elle se traduit par un balisage restreint de la structure dictionnairique grâce à l'identification de "points d'accès" au document caractérisant à la fois des indices typographiques ou de mise en page du texte balisé, pouvant ainsi en mentionner l'édition, la page, la colonne, ou préciser l'existence de fontes particulières (grandes et petites capitales, italique, gras, etc.). La figure 1 ci-dessous[1] fournit un exemple de balisage *minimal* :

> TIMBRE. s. m. Sorte de cloche ronde qui n'a point de battant en dedans, & qui est frappée en dehors par un marteau. *Le timbre d'une horloge. timbre d'un reveille-matin. le timbre de cette horloge est tres-bon.*[...] Timbrer. v. a. Terme de blason, Accompagner d'un timbre. *Timbrer une armoirie*.
> Timbrer. v. a. Terme de Pratique, Escrire au haut d'un Acte, la nature de cet acte, sa date & le sommaire de ce qu'il contient. *Timbrer des pieces*.
> On dit aussi, *Timbrer du papier, timbrer du parchemin*, pour dire, Imprimer la marque du Roy sur du papier, sur du parchemin, pour faire qu'il puisse servir aux actes de Justice.
>
> <page n="563"><col n="1">[...]<p><lc>TIMBRE</lc>. s. m. Sorte de cloche ronde qui n'a point de battant en dedans, & qui est frappée en dehors par un marteau. Le timbre d'une horloge. timbre d'un reveille-matin. le timbre de cette horloge est tres-bon.<p>
> [...]
> <sc>Timbrer</sc>. v.a. Terme de blason, Accompagner d'un timbre.
> Timbrer une armoirie.
> <p><sc>Timbrer</sc>. v.a. Terme de Pratique, Escrire en haut d'un Acte, la nature de cet acte, sa date & le sommaire de ce qu'il contient.
> Timbrer des pieces.
> <p>On dit aussi, Timbrer du papier, timbrer du parchemin, pour dire, Imprimer la marque du Roy sur du papier, sur du parchemin, pour faire qu'il puisse servir aux actes de Justice.<p>[2]

Figure 1. Exemple de balisage minimal (Dictionnaire de l'Académie 1694)

En plus de cette première forme d'indications, l'approche minimaliste permet également l'accès à certains champs informationnels de l'article balisé et ce grâce à l'utilisation de mots-clés métalinguistiques (Wooldridge, Leroy-Turcan, 1996).

---

[1] Exemple tiré de l'article de T.R. Wooldridge, "L'informatisation du Dictionnaire de l'Académie française (DAF)", dans *Actes du colloque-atelier international DictA1998* organisé par le Groupe d'Études sur l'Histoire de la Langue Française (GEHLF) et la Société Internationale d'Études Historiques et Linguistiques des Dictionnaires Anciens (SIEHLDA), Université de Limoges, 19-20 novembre 1998.

[2] Les codages <page n="563">, <col n="1">, <p>, <lc></lc>, , <sc></sc>, servent respectivement à marquer la page dans laquelle se situe l'article balisé, les colonnes occupées par l'article dans la page, la présence d'un paragraphe, les grandes capitales (Low capitales), le caractère italique de l'information codée, et les petites capitales (Small capitales).

Cette solution du balisage *minimal* contraste assez radicalement avec la solution du balisage *analytique* retenue pour l'informatisation du *Dictionnaire Universel de Furetière revu par Basnage de Bauval* (1702) (Wionet, Tutin, 1998 et 2001), dans la mesure où, en plus d'un balisage de nature typographique fin, Chantal Wionet et Agnès Tutin revendiquent un repérage exhaustif et systématique[1] des différents champs informationnels de l'article de dictionnaire, en vue de permettre des requêtes très fines de sa structure. La figure 2 proposée ci-dessous illustre le balisage *analytique* de l'article *Daguet* du *Dictionnaire Universel de Furetière revu par Basnage de Bauval* :

DAGUET. Terme de Venerie. Jeune cerf, qui est à sa premiere tête; qui pousse son premier bois.
Daguet. adv. Sourdement; en cachette. Il s'en est allé, il a tiré ses chausses daguet. Cela est bas et populaire.[2]

```
<Entry>
  <Form Type=LEMMA><Orth Rend=CAPS>DAGUET</Orth>. </Form>
  <GramGrp><Pos Type=S></Pos><Gen Type=M></Gen></Gramgrp>
    <Sense><CDomain><Lbl>Terme de</Lbl><Domain> Venerie</Domain>.</CDomain>
      <Def>Jeune cerf, qui est à sa premiere tête; qui pousse son premier bois.</Def></Sense>
    <Re><Form Type=HOMOGRAPH><Orthre Rend=SCAPS>Daguet</Orthre>.</Form>
      <GramGrp><Pos Type=ADV>adv. </Pos></GramGrp>
      <Sense><Def>Sourdement; en cachette. </Def>
      <Eg><Q>Il s'en est allé, il a tiré ses chausses <Oref Rend=IT>daguet</Oref>. </Q></Eg>
      <CUsg><Lbl>Cela est </Lbl><Usg>bas et populaire.</Usg></Cusg></Sense></Re>
</Entry>
```

Figure 2. Exemple de balisage analytique

Ces deux stratégies de rétroconversion offrent bien évidemment un rendement différent en termes de capacité d'interrogation des données rétroconverties. À ce titre, mais également parce qu'elle s'appuie sur l'utilisation d'une norme de codage[3] - le langage SGML (Standard Generalized Markup Language) - la solution du balisage *analytique* apparaît aujourd'hui comme la solution la plus adaptée pour la redécouverte des monuments de notre lexicographie grâce à l'outil informatique.

1.2. Quelques spécificités à prendre en compte

Dans la perspective de fournir une édition électronique du corpus de thèse[4] de C. REY (Rey 2004), nous avons essayé de choisir la solution de balisage la mieux adaptée à la nature encyclopédique de ce dernier. Notre volonté étant de fournir une édition électronique autorisant une interrogation la plus riche possible de la structure informationnelle, la solution du balisage *minimal* a d'emblée été écartée. Nous avons donc dans un premier temps opté pour la mise en application de la solution plus séduisante du balisage *analytique*.

Néanmoins, la nature encyclopédique du dictionnaire à rétroconvertir nous a rapidement confrontés à deux obstacles que ce type de balisage ne règle pas.

Nous avons précisé plus haut que la richesse du balisage *analytique* découlait du repérage de la totalité des champs informationnels de l'article, or, la structure "molle" de notre dictionnaire - loin d'être aussi rigide que celle des dictionnaires modernes - met en évidence l'existence de certains chevauchements d'informations. Certains articles sont en effet construits de telle manière qu'il nous est difficile de savoir où commence et où finit un champ particulier. C'est notamment ce que

---

[1] Lorsqu'un champ informationnel "stable" n'existe pas, il est tout de même procédé à un balisage spécifiant à l'utilisateur que ce champ est effectivement absent.

[2] Sans trop détailler l'exemple fourni, nous pouvons mentionner le repérage des champs informationnels principaux tels que <Form> indiquant si l'entrée ou la sous-entrée est un lemme ou un homographe, la marque de domaine <Domain>, l'information grammaticale <GramGrp>, la définition <Def>, et des champs secondaires comme les remarques sur la typographie des entrées, des sousentrées, ou des références, <Orth Rend=CAPS>, <Orthre Rend=SCAPS>, <Oref Rend=IT>, les marques d'usage <Usg>, les particules introduisant un champ, nommées libellés <Lbl>, etc.

[3] La solution du balisage *minimal* s'appuie elle sur des balisages propriétaires, à savoir des jeux de balises ne faisant pas l'objet de recommandations par un organisme de standardisation.

[4] Ce corpus va être présenté dans la seconde partie de cette communication.

nous avons essayé d'illustrer à travers les découpages multiples envisageables pour l'énoncé suivant tiré de l'article *Sous-entendu* du *Dictionnaire françois* (1680) de César-Pierre Richelet :

> 1- **(** C'est une figure de Grammaire**) [**qui consiste à n'exprimer point, par élégance, un ou plusieurs mots.**]**
> 2- **[**C'est une figure **(**de Grammaire**)** qui consiste à n'exprimer point, par élégance, un ou plusieurs mots.**]**
> 3- **[**C'est une figure**] (**de Grammaire**) [**qui consiste à n'exprimer point, par élégance, un ou plusieurs mots.**]**

Dans cet exemple précis, l'identification de la marque de domaine grammatical - délimitée par les parenthèses - et de la définition - mise entre crochets - ne peut se faire sans que l'arbitraire du lexicographe procédant au balisage ne se manifeste. Une telle difficulté, déjà présente pour des dictionnaires de langue et exacerbée pour les dictionnaires encyclopédiques, réside dans notre volonté de vouloir imbriquer chaque portion textuelle au sein de "boîtes" pré-étiquetées et distinctes les unes des autres.
En lien étroit avec ce premier obstacle, il semblerait par ailleurs que la nature encyclopédique de notre dictionnaire génère également une multiplication des informations susceptibles d'être mises en valeur dans le but d'une investigation lexicographique informatique plus fine.
À titre d'exemples, il peut en effet paraître intéressant de repérer les informations issues du développement encyclopédique que ne fournit pas le dictionnaire de langue. À un autre niveau d'analyse, il peut également s'avérer important de baliser les titres d'ouvrages ou les noms propres mentionnés dans le corps de l'article. Comment peut alors se faire le repérage de ces données sans que de nouveaux chevauchements informationnels ne soient générés ?
Les réflexions que nous avons été amenés à conduire pour mettre au point la rétroconversion de notre corpus se sont concrétisées par l'émergence d'une nouvelle solution de balisage que nous allons détailler ci-dessous : le balisage *souple* ou *flottant*.

2. Déconstruire positivement le texte ancien

La solution du balisage *souple* ou *flottant*, en plus de constituer une alternative intéressante aux solutions déjà existantes, participe selon nous à la "déconstruction positive du texte ancien". Pour illustrer les particularités de ce balisage, nous nous proposons d'évoquer l'exemple de la "déconstruction informatique positive" que nous avons fait subir à notre corpus de travail : les articles de Grammaire de l'*Encyclopédie Méthodique*.
Conçue comme une édition corrigée, remaniée et augmentée de la célèbre *Encyclopédie* ou *Dictionnaire raisonné des arts et des sciences* (1751-1777) de Diderot et d'Alembert, l'*Encyclopédie Méthodique* (1782-1832) publiée par Charles-Joseph Panckoucke constitue l'un de nos monuments lexicographiques du siècle des Lumières. À plusieurs égards décisive pour la traduction de la mutation épistémologique s'étant imposée entre le milieu et la fin du XVIII$_{e}$ siècle, la "Méthodique" mérite ainsi de sortir de l'ombre imposante de la première encyclopédie. C'est à ce titre qu'après avoir fourni une réflexion d'historien de la langue sur la nature des connaissances véhiculées sur les sons du français au sein du dictionnaire *Grammaire & Littérature* (1782-1786) - l'un des trente-neuf dictionnaires de matière de cette encyclopédie – nous nous sommes penchés sur l'élaboration d'une méthode d'informatisation qui puisse restituer toute la richesse et l'originalité de cet ouvrage.

2.1. Pourquoi parler de déconstruction ?

Ainsi que nous l'avons sommairement évoqué plus haut, notre conception de déconstruction du document est issue de l'initiative présente au sein du balisage *analytique* de Wionet et Tutin.
En effet, à la différence du balisage *minimal* qui ne propose qu'un accès réduit au contenu de l'article, le balisage *analytique* propose de poser au sein de la microstructure de ce dernier des jalons issus d'une norme de codage, le langage SGML. Notre solution propose le même "éclatement" informationnel mais privilégie la norme de codage XML (eXtensible Markup Language), standard incontesté de codage et de diffusion des documents électroniques.
L'éclatement occasionné par un balisage *souple* ou *flottant* se distingue néanmoins de celui obtenu grâce à un balisage *analytique* par le fait qu'il s'affranchit comme nous l'avons déjà dit plus haut, de l'idée de faire coïncider chaque portion de texte avec un champ informationnel préétabli. Toute la flexibilité et la souplesse de notre balisage vient d'un dégroupement des données de

l'article en deux grands blocs : un bloc <entree> qui renferme le lemme (ou ses formes fléchies et dérivées) ainsi que l'information grammaticale, et un bloc <corps> qui renferme la totalité des autres informations de l'article. C'est précisément au niveau du bloc <corps> que les problèmes d'identification et de découpage des éléments sont les plus présents et que le balisage *souple* ou *flottant* prend tout son sens.
Ainsi que l'illustre la figure 3 ci-dessous représentant la forme balisée de l'article *Prosonomasie*[1], les données appartenant à ce bloc peuvent flotter à l'intérieur de cet espace en étant ou non identifiées comme appartenant à un champ informationnel clairement défini.

```
<ARTICLE>
      <STATUT TYPE="DIFFERENT"/>
      <ENTREE TYPE="EP">
            <FORME>PROSONOMASIE</FORME>,
                  <INFORMATION_GRAMMATICALE TYPE="SUBSTANTIF
                  FEMININ">
                        <PARTIE_DU_DISCOURS TYPE="SUBSTANTIF">s.
                        </PARTIE_DU_DISCOURS>
                        <GENRE TYPE="FEMININ">f.</GENRE>
                  </INFORMATION_GRAMMATICALE>
      </ENTREE>
      <CORPS>
            <EXTRA TYPE="AJOUT">(<ETYMOLOGIE><LANGUE
            TYPE="GREC">Προσωνομασία</LANGUE>, du verbe
            <LANGUE TYPE="GREC">προσωνομάζω </LANGUE>,
            <LANGUE TYPE="LATIN">insuper nomino
            </LANGUE>.</ETYMOLOGIE> <DEFINITION>C'est un autre nom de la
            figure appelée ordinairement Paronomase).</DEFINITION>
            <SIGNATURE TYPE="BEAUZEE">(M.BEAUZÉE.)</SIGNATURE>
            </EXTRA><DISCOURS_ENCYCLOPEDIQUE>Figure de Rhétorique, par
            laquelle on fait allusion à la ressemblance du son qui se trouve entre différents
            noms ou différents mots, comme dans ces phrases : <LANGUE
            TYPE="LATIN">Is verè consul est qui Reipublicæ saluti consulit ; Quum
            lectum petis de lethocogita </LANGUE>. Voyez <REFERENCE
            TYPE="VEDETTE">PARONOMASE</REFERENCE>.
            </DISCOURS_ENCYCLOPEDIQUE>
            <SIGNATURE TYPE="ANONYME">(ANONYME.)</SIGNATURE>
      </CORPS>
</ARTICLE>
```

Figure 3. Exemple de balisage Souple ou Flottant

À l'intérieur de ce bloc, des données peuvent également être identifiées comme appartenant à un champ particulier, tout en étant déjà incluses dans un autre champ. Ceci donne l'impression que ces éléments constituent des portions de texte flottantes.
Une autre particularité de notre balisage est de dissocier le balisage *logique*, à savoir le repérage des champs informationnels, du balisage *physique* qui considère les critères typographiques du document informatisé. Par le biais d'une feuille de style associée au document .xml contenant le texte balisé, il est alors possible de restituer l'aspect physique du texte rétroconverti. Il est par ailleurs intéressant de mentionner que le balisage physique du document permet, ainsi que cela a pu être mis en application pour le *Tlfi* (Dendien, Pierrel 2003), d'effectuer un balisage logique semi-automatique du document.

2.2. L'apport "positif" de l'Informatique

L'aspect "positif" de la déconstruction que nous proposons grâce au balisage *souple* ou *flottant* se trouve quant à lui illustré par les possibilités d'interrogation du document qui découlent de ce processus de balisage.
Dans une entreprise telle que la rétroconversion d'un dictionnaire, les capacités d'interrogation des données et d'investigation au coeur du texte vont entièrement dépendre de la richesse du balisage appliqué. Pour notre entreprise d'informatisation des articles du dictionnaire *Grammaire & Littérature* de l'*Encyclopédie Méthodique*, la puissance d'interrogation des données repose sur un

[1] La *Prosonomasie* est une figure de Rhétorique par laquelle on fait allusion à la ressemblance du son qui se trouve entre différents noms ou différents mots.

savant mélange de technologies : 1) le choix d'une Définition de Type de Document (DTD) à la fois rigoureuse, souple et riche, et 2) le choix des outils de développement.
La DTD TEI existant pour les dictionnaires n'étant pas pleinement satisfaisante pour la rétroconversion des dictionnaires anciens (Cf. Rey 2004), et qui plus est pour des dictionnaires anciens de nature encyclopédique, nous avons élaboré notre propre DTD.
Sans pour autant détailler celle-ci, nous pouvons mentionner le fait que cette dernière concilie à la fois le repérage d'éléments traditionnellement identifiés dans la microstructure classique d'un dictionnaire (information grammaticale, information étymologique, définition, etc.) et des éléments supplémentaires (titres d'ouvrages, noms de personnes, noms de personnages, etc.).
Afin d'exploiter au mieux cette DTD, nous avons eu recours à un ensemble d'outils de développement à la fois novateurs, gratuits et permettant de livrer un exécutable.
Notre choix s'est porté sur une Interface de Programmation (API, Application Programming Interface) et plus précisément sur la représentation hiérarchique en mémoire du document XML sous forme d'arbre fournie par l'API DOM (Document Object Model).
Parmi les langages permettant d'exploiter l'API DOM, nous avons retenu le langage de programmation C++ en association avec les librairies QT.
De tous ces choix méthodologiques a pu émerger un outil d'interrogation de notre corpus de travail : le logiciel CorpXML[1].
Conformément à notre souci de proposer un outil qui ne soit pas spécifique à notre seul corpus, le logiciel CorpXML ne s'appuie pas directement sur notre DTD et reconstitue de manière dynamique la structure de tous les documents qui lui sont proposés en lecture.
Deux grands types de recherche sont intégrés au logiciel CorpXML. Une recherche dite "simple" permet d'effectuer une interrogation sur le corpus étudié à partir de mots-clés.
La figure 4 fournit ainsi un exemple de recherche visant à identifier tous les articles comportant à la fois le mot "air" et le mot "nasal" :

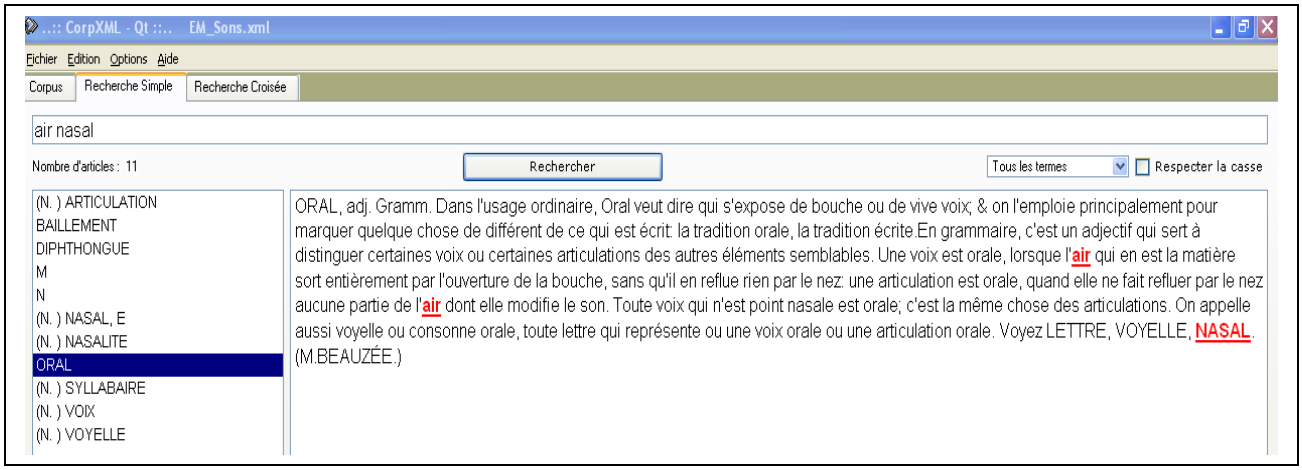


Figure 4. Recherche simple du logiciel CorpXML.

Dans la colonne de gauche s'affiche la liste des articles contenant les mots recherchés, tandis que la partie droite de l'écran permet de lire l'article sélectionné par l'utilisateur au sein de cette liste. Les mots ayant fait l'objet de la recherche sont quant à eux mis en évidence par un procédé visuel combinant un soulignement et une coloration en rouge.
Le second type de recherche, la recherche dite "croisée", repose plus particulièrement sur les balises XML qui jalonnent le texte. Elle permet de croiser de manière assez fine plusieurs critères de recherche.
En ce qui concerne notre corpus d'étude, il est par exemple possible, ainsi que l'illustre la figure 5 ci-dessous, d'obtenir très facilement l'ensemble des articles rétroconvertis constituant des **Substantifs** appartenant au domaine de la **Grammaire** (noté "Gramm") et comportant des portions de texte en **Latin** :

[1] Cet outil gratuit est téléchargeable à l'adresse internet suivante : http://www.up.univmrs.fr/delic/perso/rey/methodique/index.htm

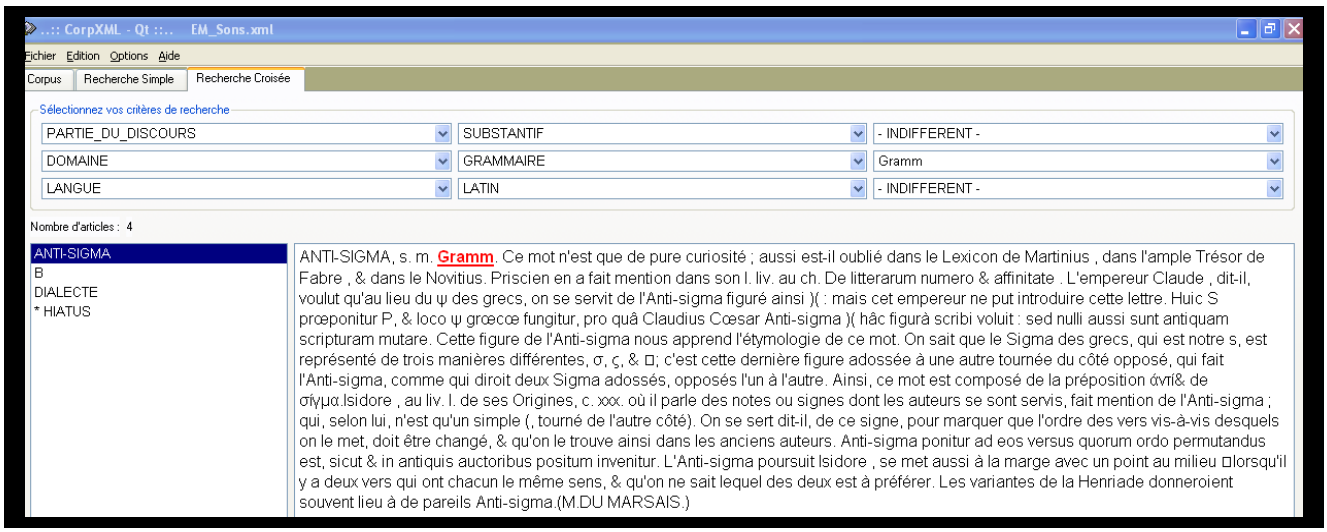


Figure 5. Recherche croisée du logiciel CorpXML.

La colonne de gauche de cette figure fournit la liste des articles correspondant à nos critères après l'application de nos filtres successifs. Dans la partie droite de l'interface, on visualise l'article sélectionné dans la liste obtenue et les portions de texte ayant éventuellement été spécifiées par les critères de recherche sont mises en valeur par le procédé graphique déjà mentionné plus haut.

Conclusion

Ainsi que nous l'avons montré ici, la solution du balisage *souple* ou *flottant* pour la rétroconversion des dictionnaires anciens constitue le point d'intersection de deux visions de l'informatisation des données.

Dans un premier temps, cette solution constitue en effet une réponse à la volonté du lexicographe de ne pas "corrompre" la structure "molle" du dictionnaire encyclopédique ancien en cherchant à tout prix à retrouver au sein de ce dernier la rigueur structurelle de nos dictionnaires modernes.

Dans un deuxième temps, cette même solution répond au souci de l'informaticien d'utiliser une technologie permettant à la fois de respecter la souplesse du texte informatisé et d'en autoriser une interrogation à la fois riche et en conformité avec le courant actuel d'informatisation des documents.

Testé avec succès sur un corpus issu d'une encyclopédie du siècle des Lumières, le balisage *souple* ou *flottant* semble constituer une alternative intéressante aux solutions d'informatisation des dictionnaires anciens déjà existantes, et mérite certainement d'être mis en application sur d'autres ouvrages.

BIBLIOGRAPHIE

BEAUZÉE, N. & MARMONTEL, J-F. (1782-1784-1786). *Encyclopédie Méthodique. Grammaire & Littérature*, Paris (chez Panckoucke), Liège (chez Plomteux). 3 vol.

CARON, P., DAGENAIS, L., GONFROY, G. 1992. Le programme d'informatisation du Dictionnaire critique de la langue française de l'abbé Jean-François Féraud (1787), in T.R. Wooldridge (éd.), *Historical Dictionary Databases*, *CCH Working Papers*, 2, pp. 87-103.

DENDIEN, J. & PIERREL, J-M. 2003. Le trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence, *TAL*, Volume 44 - n°2/2003, 28 p.

Le Dictionnaire de l'Académie française : histoire et nuances de la langue française (1694- 1935), (2000), éditions Redon.

PANCKOUCKE, Ch-J. (1782-1832). *Encyclopédie méthodique ou par ordre de matières par une société de gens de lettres, de savants et d'artistes; précédée d'un Vocabulaire universel, servant de Table pour tout l'Ouvrage, ornée des Portraits de MM. Diderot et d'Alembert, premiers Éditeurs de l'Encyclopédie*, Paris, Panckoucke.

REY, C. 2004. *Analyse et informatisation des articles traitant de l'étude des sons dans le dictionnaire Grammaire & Littérature de Nicolas Beauzée et Jean-François Marmontel, issu de l'Encyclopédie Méthodique*, Thèse de doctorat, Aix-en-Provence.

REY, C. & ZAOUI, C. 2004. Le balisage XML "ciblé" : une nouvelle approche dans l'informatisation des corpus, in *Actes de la conférence internationale sur la Fouille de Texte* (CIFT'04), dans le cadre de la semaine du Document Numérique, La Rochelle, 22-24 juin 2004, pp. 121-133.
RICHELET, C-P. 1680. *Dictionnaire françois, Dictionnaires des XVIe et XVIIe siècles*, 1998, cd-rom pc, version 1.0, Champion électronique.
VÉRONIS, J., & IDE, N. 1996. Encodage des dictionnaires électroniques : problèmes et propositions de la TEI, in D. Piotrowsky (éd.), *Lexicographie et informatique - Autour de l'informatisation du Trésor de la Langue Française*, Actes du Colloque International de Nancy (29, 30, 31 mai 1995), Paris, Didier Erudition, pp. 239-261.
WIONET, C. & TUTIN A. 2001. *Pour informatiser le Dictionnaire universel de Basnage (1702) et de Trévoux (1704) Approche théorique et pratique*, Paris, Honoré Champion.
WIONET C., TUTIN A. 1998. Informatisation du Dictionnaire Universel de Furetière revu par Basnage de Bauval (1702) : premier bilan, in *Actes du colloque-atelier international DictA1998* organisé par le Groupe d'Études sur l'Histoire de la Langue Française (GEHLF) et la Société Internationale d'Études Historiques et Linguistiques des Dictionnaires Anciens (SIEHLDA), Université de Limoges, 19-20 novembre 1998.
WOOLDRIDGE, T.R. 1998. L'informatisation du Dictionnaire de l'Académie française (DAF), in *Actes du colloque-atelier international DictA1998* organisé par le Groupe d'Études sur l'Histoire de la Langue Française (GEHLF) et la Société Internationale d'Études Historiques et Linguistiques des Dictionnaires Anciens (SIEHLDA), Université de Limoges, 19-20 novembre 1998.
WOOLDRIDGE, T.R. & LEROY-TURCAN I. 1996. Les mots-clefs métalinguistiques comme outil d'interrogation structurante des dictionnaires anciens, in A. Clas, P. Thoiron & H. Béjoint (éds.), *Lexicomatique et dictionnairiques*, Beyrouth, FMA & Montréal, AUPELFUREF, pp. 307-16.
WOOLDRIDGE, T.R. 1994. Projet d'informatisation du Dictionnaire de l'Académie (1694-1935), in B. Quemada & J. Pruvost (éds.), Actes du Colloque international *Le Dictionnaire de l'Académie française et la lexicographie institutionnelle européenne*, Institut de France, novembre 1994, Paris, Champion, pp. 309-20.
WORLD WIDE WEB CONSORTIUM. Extensible Markup Language (XML) : http://www.w3.org/XML/.

ÉCRIRE EN CRITIQUE : EXPLORATION MORPHO-SYNTAXIQUE SUR CORPUS

Driss ABLALI
Université de Franche-Comté (LASELDI)

« C'est le langage qui parle, ce n'est pas l'auteur ». Mallarmé

Introduction

Différentes études ont mis en évidence des variations systématiques des catégories morphosyntaxiques pour caractériser et discriminer les discours et les genres[1]. Cet article explore, à travers la comparaison de deux corpus de critique, critique littéraire et critique journalistique, les indices de divergence et de convergence entre discours que permettent d'observer les méthodes exploratoires de données textuelles. À partir d'une réflexion sur le genre, on cherche à mettre en évidence les traits discriminants permettant d'identifier le genre de la critique loin du carcan littéraire où il a longtemps été cantonné : existe-t-il une identité morphosyntaxique du genre de la critique ? Existe-t-il un style, des catégories grammaticales, des signes de ponctuation qui caractérisent l'article du journaliste et celui du littéraire et qui définissent une posture discursive originale ? Est-ce que ce qui fonde un genre est de l'ordre de l'autoproclamation de ce genre, comme cela a pu être le cas dans les théories littéraires, ou de l'ordre de la situation des discours ? Notre propos n'est évidemment pas de réhabiliter le critique, nous interrogerons le corpus pour découvrir les spécificités du discours dans lequel le même genre, à savoir la critique, s'inscrit, le discours des études littéraires et le discours journalistique. Il est temps en effet de remettre en cause la démarche littéraire qui, en disqualifiant la notion de discours, fait des genres le niveau le plus englobant de sa typologie.

Dans la perspective de la sémantique interprétative de F. Rastier[2], nous distinguons quatre niveaux hiérarchiques supérieurs : les *discours* (ex. juridique *vs* littéraire *vs* essayiste *vs* scientifique), les *champs génériques* (ex. à l'intérieur du discours littéraire : théâtre, poésie, genres narratifs), les *genres* proprement dits (ex. comédie, roman « sérieux », roman policier, nouvelles, contes, mémoires et récits de voyage), les *sous-genres* (ex. roman par lettres). Au niveau inférieur de la classification, nous trouvons les textes de même genre et d'un même auteur.

En questionnant les modalités grammaticales et syntaxiques du genre de la critique, le but est d'identifier un régime singulier du critique littéraire et du critique journaliste. Quelles en sont les caractéristiques et sous quelles formes se manifeste-t-il ? Interroger le même genre au sein de deux discours différents, c'est donc analyser les liens qu'il tisse entre morphologie et syntaxe.

Notre travail portant sur le genre de l'article de critique, il importait, dans la perspective d'une étude contrastive en corpus, de pouvoir à la fois opposer la critique littéraire à la critique journalistique et en même temps de pouvoir montrer que le genre n'est pas envisageable sans discours. Nous explorerons d'abord la longueur du mot, facteur déterminant quant au rythme du texte. Un autre facteur révélateur du style d'un genre est la longueur de la phrase et le nombre de mots par proposition. Nous nous intéresserons par la suite à la segmentation à l'intérieur de la phrase en analysant les signes qui la représentent : la virgule, le point-virgule, les deux-points ainsi que les signes de parenthèse. Nous terminerons sur la question des catégories grammaticales, pronoms personnels et temps verbaux.

Les paramètres du corpus

Le corpus compte 5 214 615 d'occurrences. Il comprend uniquement des articles intégraux et non des extraits. Il se répartit sur deux discours : littéraire et journalistique. 2 041 875 en critique littéraire, contre 3 116 740 en critique journalistique. Le premier comprend des articles tirés de revues françaises ou francophones relevant tous du domaine de la critique littéraire. Il contient 261 articles, publiés entre 1980 et 2004, extraits de 11 revues, comme *RITM, SEMEN, TEXTE, CAHIERS DE NARRATOLOGIE, CHAMPS DE SIGNE, ERITA, LECTURE LITTERAIRE, LOXIAS*. Le second comporte les articles de deux quotidiens français *LIBERATION* et *LE MONDE*, publiés

[1] Cf., entre autres, Rastier & Malrieu 2001, Beauvisage 2001, Loiseau & Poudat & Ablali 2006.
[2] Cf. Rastier 2001.

entre 2002 et 2003 dans la rubrique « critique », soit 1950 articles. Quantitativement, le corpus se présente ainsi :

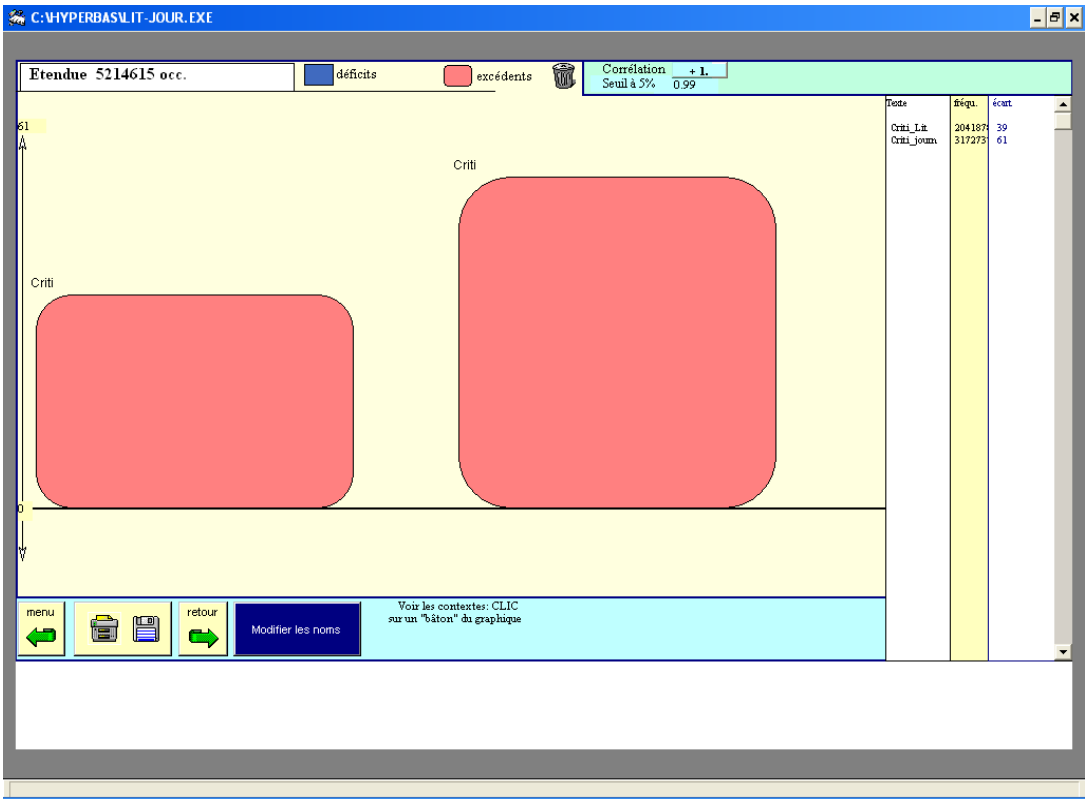


Fig. 1 : Étendue relative du corpus « Critique »

Segmentation de la phrase critique

Afin d'opposer critique littéraire et critique journalistique, une première démarche consiste à comparer l'écart entre les deux pour chaque variable, en se basant sur les moyennes des valeurs des deux discours constituant les deux ensembles. Ici il s'agit de faire une exploration morphosyntaxique du corpus pour voir la manière dont sont agencées les phrases et les propositions, ainsi que l'examen de la distribution quantitative des mots à l'intérieur des segments, de la longueur des mots, tous rendus possibles par l'application des logiciels adaptés. Les données morphosyntaxiques dont nous disposons pour mener cette étude ont été produites par la société Synapse, à l'aide du logiciel Cordial.

Pour la longueur du mot, nous avons remarqué que le nombre moyen de lettres par mot est de : 4,65 pour le discours journalistique, et 4,89 pour le discours littéraire. Rien d'étonnant : la concision, c'est l'ergonomie du journalisme, c'est obtenir le même résultat informatif en moins de mots, en moins de phrases, moyennant moins de "bruit". « Dans le cas du français moderne, écrit C. Muller, le recours à un vocabulaire technique, savant ou simplement recherché augmente la proportion des mots longs, faisant ainsi monter notre indice ; un style familier ou relâché, sur le plan du lexique, agit évidemment en sens inverse. D'autre part une syntaxe soignée tend à réduire la densité du texte en mots grammaticaux, donc en mots courts, éléments que la langue courante multiplie » (1979 :153). Ainsi, la longueur du mot nous permet de confirmer l'opposition des deux discours, les mots longs se trouvant plutôt dans le discours littéraire, et les mots courts dans le discours journalistique. Et cette opposition générique se manifeste aussi au niveau de la proposition : pour le discours journalistique, le nombre moyen de mots par proposition est de 8,86, alors que pour le discours littéraire, il est de 10,48. Un autre facteur, plus souvent étudié que celui de la longueur du mot, est la longueur de la phrase, son unité supérieure, considérée comme un critère déterminant dans l'analyse stylistique et dans l'étude du rythme des textes. Comme critère stylistique fortement caractéristique d'un auteur, d'un genre ou d'un discours, elle permet ainsi d'appréhender des variations intéressantes, notamment d'un point de vue diachronique, pour étudier l'évolution de la longueur de la phrase dans plusieurs œuvres successives d'un même auteur, ou d'un point de vue contrastif, soit à l'intérieur d'un même discours, opposer par exemple

le portrait au reportage au sein du discours journalistique, soit à partir de discours différents, comme nous le faisons ici, opposer le discours journalistique au discours littéraire. La longueur moyenne de la phrase s'obtient en divisant le nombre d'occurrences du corpus par le nombre de ponctuations fortes. Lorsque l'on compare la longueur des deux discours, encore une fois le discours littéraire dépasse le journalistique. Ce dernier se caractérise par un nombre important de phrases courtes. Il enregistre une moyenne de 18,63 mots par phrase contre 22,02 pour l'article littéraire. Une différence de 3,39 est énorme, surtout lorsqu'il s'agit du même genre. C'est encore une fois la preuve que les caractéristiques stylistiques et syntaxiques ne dépendent pas du genre dans lequel le texte s'inscrit mais de la situation des discours. « À chaque discours, on peut faire correspondre un système ou *symmorie* générique. Chaque groupe de pratiques sociales correspondant à un discours se divise en activités spécifiques », nous dit F. Rastier, (2001 : 4), qui fait ici le postulat que, par exemple, non seulement le discours juridique se différencie du discours médical ou du discours littéraire, mais qu'il faudrait trouver les variables qui discriminent de façon préférentielle chaque niveau de la hiérarchie : discours, champs génériques, genres, sous-genres. Ces variables, nous les retrouvons aussi au niveau de la proposition. Si l'on regarde l'histogramme ci-dessous, on voit sans équivoque que le discours journalistique est plus riche en propositions indépendantes, 88,23% contre 84,44 pour le discours littéraire. Car l'écriture journalistique requiert des règles et codes spécifiques qui structurent, organisent et codifient son contenu pour optimiser sa transmission au lecteur : comme l'espace est compté, le lecteur est présumé pressé, le style doit donc être concis. Et puisque le lecteur n'a pas de temps pour le verbiage, la chasse aux propositions relatives (1,86% *vs* 2,60%) et aux propositions subordonnées (8,63% *vs* 11,11%) devient une priorité. L'histogramme ci-dessous présentant les variations de la critique littéraire par rapport à la critique journalistique, montre l'ensemble des variables des deux discours :

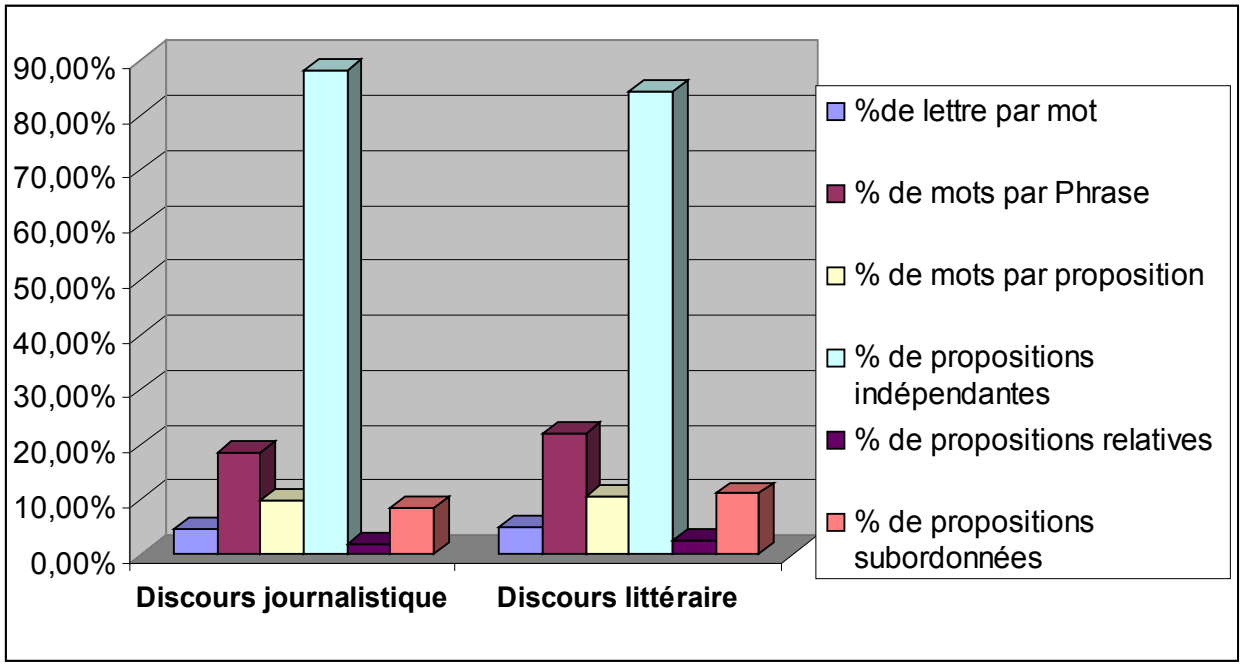


Fig. 2 : Segmentation de la phrase dans les deux discours

Toutefois, la longueur des mots et des phrases ne suffit pas à elle seule à caractériser les textes, ce qui nous amène à prendre en compte aussi les marques de ponctuation. Pour le point, principal marqueur de fin de phrase, nous observons que sa fréquence est plus élevée dans le discours journalistique. Plus de phrases courtes engendre automatiquement plus de points. Le pourcentage de points par rapport à l'ensemble des ponctuations est de 36,56 pour l'article journalistique, et de 22,34 pour l'article littéraire. Quant à la virgule, son effectif suit le cours logique des choses : le texte qui développe l'effectif le plus important en points fait usage de moins de virgules par rapport à l'autre discours. Dans notre corpus, la fréquence des phrases courtes est importante dans le discours journalistique, l'effectif de la virgule par rapport à l'ensemble des signes de ponctuation est de 45,88, alors qu'il est de 52,7 pour le discours littéraire. Or pour les autres marques de

ponctuation, toujours fortes, comme le point d'exclamation et le point d'interrogation, c'est dans le discours littéraire que nous trouvons les plus hautes fréquences, comme on le voit dans le graphique ci-dessous. L'article littéraire fait un usage beaucoup plus important du point d'exclamation que le discours journalistique. Comme le dit Culioli, le point d'exclamation entend signifier « le haut degré d'une propriété ». Le lecteur est lui-même appelé à le prendre en charge. Simplement il faut rappeler aussi que cette haute fréquence est liée à des fins intertextuelles. Le critique littéraire appuie sa thèse en recourant à des extraits des textes analysés. C'est ce qui justifie aussi la présence massive d'un autre signe de ponctuation, comme le tiret, lié à la présence de dialogue pour illustrer des propos, et qui est pratiquement absent du discours journalistique. Quant au point d'interrogation, il est dominant dans le discours littéraire, pas seulement pour la question de l'intertexte, mais surtout pour des raisons liées à la spécificité du genre de la critique littéraire, dont la thèse constitue le parangon. Cette constatation n'a rien d'étonnant, le critique écrit pour répondre à une problématique, développe des hypothèses, qu'il construit lui-même dans l'introduction de son article avec des interrogations soit directes, soit indirectes. Or dans l'article journalistique, on prononce des jugements sur les livres et les films qui paraissent, on tranche entre les bons et les mauvais sans en problématiser le contenu. On trouve aussi d'autres signes de ponctuation, au sein de l'article littéraire, qui semblent ainsi se rattacher à un pôle plus « scientifique », discriminé par une présence plus importante des crochets et accolades ou l'usage des parenthèses.
Le graphique ci-dessous, dont les résultats sont réalisés avec Cordial-Synapse, présente les caractéristiques des signes de ponctuation des deux discours :

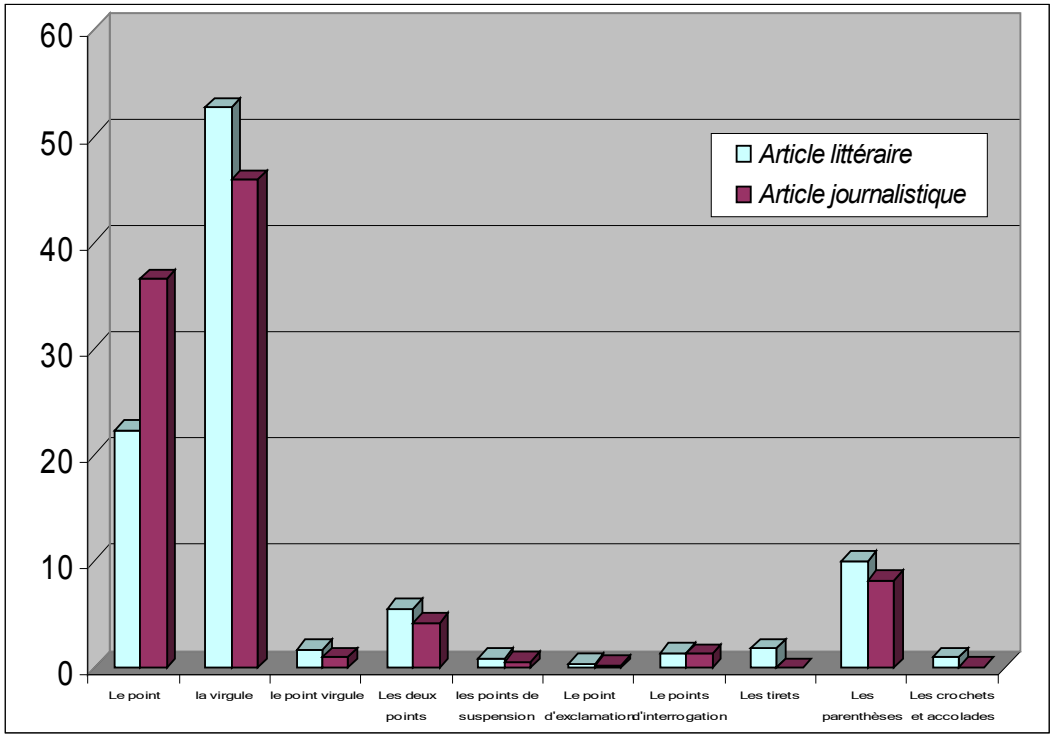


Fig. 3 : La distribution des signes de ponctuation (effectifs)

Sur la base de ces oppositions, on peut d'ores et déjà tenter d'esquisser au niveau phrastique un portrait-type de l'article journalistique par rapport à l'article littéraire. On note que les deux discours sont significativement distincts sur le plan morphosyntaxique : le discours journalistique exploite différentes techniques avec une dominance des mots courts, des phrases concises rejetant la coordination et la subordination, une diminution des virgules au profit des points, sont le signe d'une écriture plus incisive, à laquelle il faut sans doute rattacher des catégories grammaticales, des temps verbaux et pronoms personnels plus discriminants.

Catégories grammaticales, temps et personnes

« C'est bien en effet dans la distribution des catégories grammaticales et principalement des classes nominale et verbale, que se manifeste la distinction des styles, des genres et des écrivains », écrit E. Brunet, (1983 : 836), nous rajoutons des discours aussi, car pas de genre sans discours. Il n'y pas, par exemple, un discours épistolaire, il y a un genre épistolaire, comme la lettre, entre autres, qui prend des manifestations discursives différentes, selon qu'elle s'insère dans le discours des écrits ordinaires, comme le cas des lettres adressées par les allocataires à leur caisse d'allocations familiales, ou selon qu'elle est écrite par Valmont à Cécile dans *Les Liaisons Dangereuses*, comme c'est le cas dans le discours littéraire. « Investigating this question, écrivent Douglas Biber et ses collègues, can help to understand how different varieties exploit the grammatical categories of words available to them… Using the corpus, we can analyze the distribution and function of these different categories of words and study the part that they play in fulfilling the communicative function of different registers » (Biber, Conrad, Reppen 1998 : 57). La question à laquelle nous voudrions répondre dans ce dernier point s'inscrit dans le sillage de la citation de Biber : quelles sont les contraintes interprétatives mésosémantiques liées à la structure textuelle interne au genre de la critique ? Existe-t-il des signatures journalistique et littéraire en termes de catégories grammaticales, de pronoms personnels et de temps verbaux qui corroborent ce qui a déjà été discriminant ci-dessus ?

Commençons par l'étude des parties du discours. Quelles sont les particularités grammaticales qui contribuent au style de l'écriture journalistique et littéraire ? Le tableau ci-dessous présente la distribution des parties du discours des deux discours :

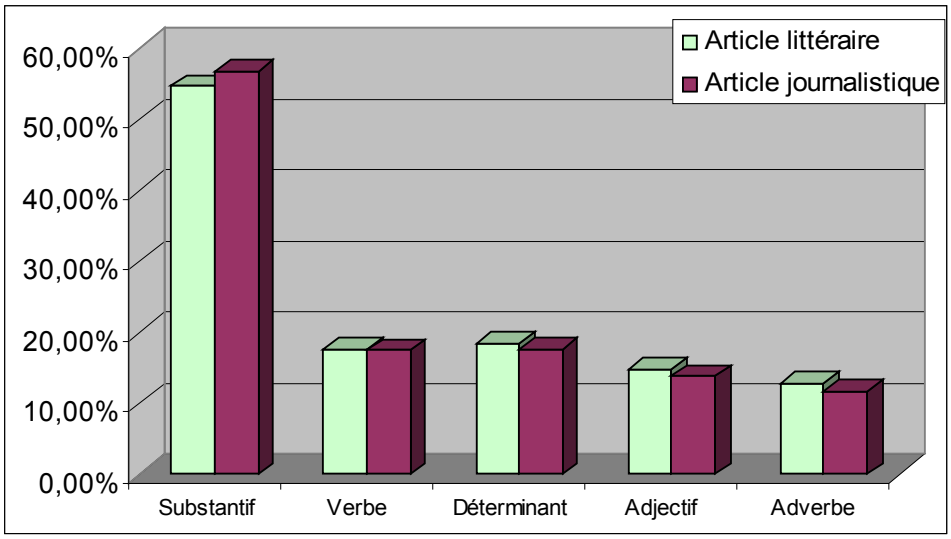


Fig. 4 : Classement hiérarchique de cinq catégories grammaticales

Première remarque : les deux discours sont riches en substantifs, une des caractéristiques des corpus de langue française. Pour le discours journalistique, la structure syntaxique la plus assimilable, la mieux comprise par le lecteur, est la structure habituelle : Sujet-Verbe-Objet. Cela correspond, comme on l'a vu ci-dessus, à la concision de l'idée. Le substantif est la catégorie grammaticale la plus employée dans les deux discours, avec une moyenne plus grande pour le discours journalistique : 56,87 contre 54,61. Pour les trois autres catégories grammaticales, les valeurs associées au discours journalistique sont plus faibles. Nous notons également la corrélation entre la fréquence du substantif et la fréquence des phrases courtes dans le discours journalistique, car, comme on le voit, le substantif journalistique est plus solitaire que son homologue littéraire. Les adjectifs et les adverbes sont plus dominants dans le discours littéraire : 14,83 contre 13,89 pour les premiers, et 12,93 contre 11,71 pour les seconds. Quant aux pronoms, notamment les pronoms personnels, car ce sont eux qui marquent la grande majorité des occurrences par rapport à l'ensemble des pronoms, 58,68 pour le discours littéraire, 60,89 pour le discours journalistique. Au niveau de la première personne aussi bien du singulier que du pluriel, l'article littéraire arrive en première position, contrairement à la troisième personne, qui domine dans le discours journalistique aussi bien avec « il » qu'avec le « on » :

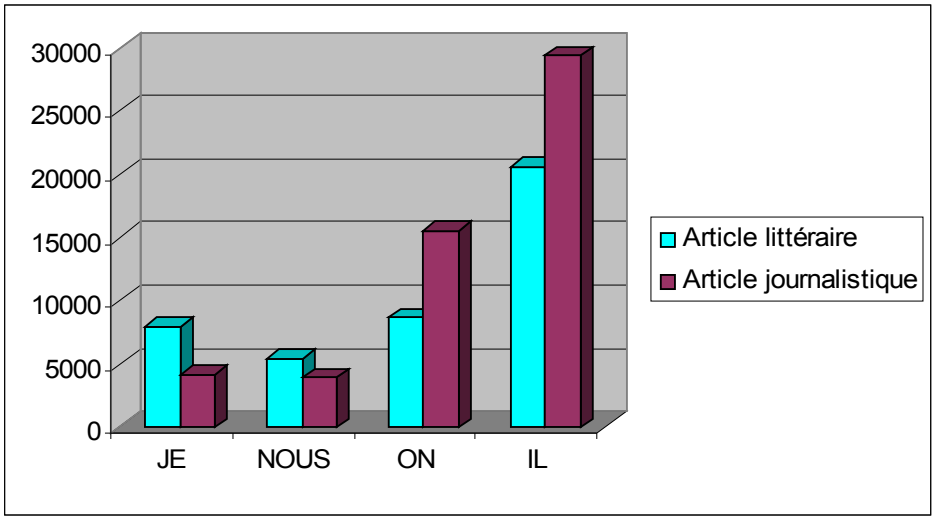


Fig. 5 : La distribution des pronoms dans le corpus

Dans cet histogramme on voit bien que le « il » est le pronom personnel le plus dominant. Une seule explication se cache derrière cette haute fréquence : dans le genre de la critique, il est question indifféremment de textes et d'auteurs, c'est-à-dire d'objets ou d'êtres absents, ceux que le critique analyse en ayant recours à la 3ème personne, laquelle assure une reprise sémantique d'un précédent. Quant au « on » son statut est plus problématique. Le « on » par son caractère malléable et sujet à changer, permet d'alterner les points de vue. Il peut désigner l'auteur ou le journaliste comme responsable de leur propos, comme il peut désigner une communauté non déterminée de gens plus ou moins compétents dans le domaine concerné. Le « on » permet de prendre ces distances par rapport à l'événement ou à la problématique développés, de même qu'il permet d'exprimer explicitement une idée quelconque. Or lorsqu'on regarde le même histogramme en se concentrant sur la première personne, aussi bien du singulier que du pluriel, en regardant leur fréquence parmi l'ensemble des pronoms personnels de notre corpus, la différence entre les deux discours est plus aisément observable :

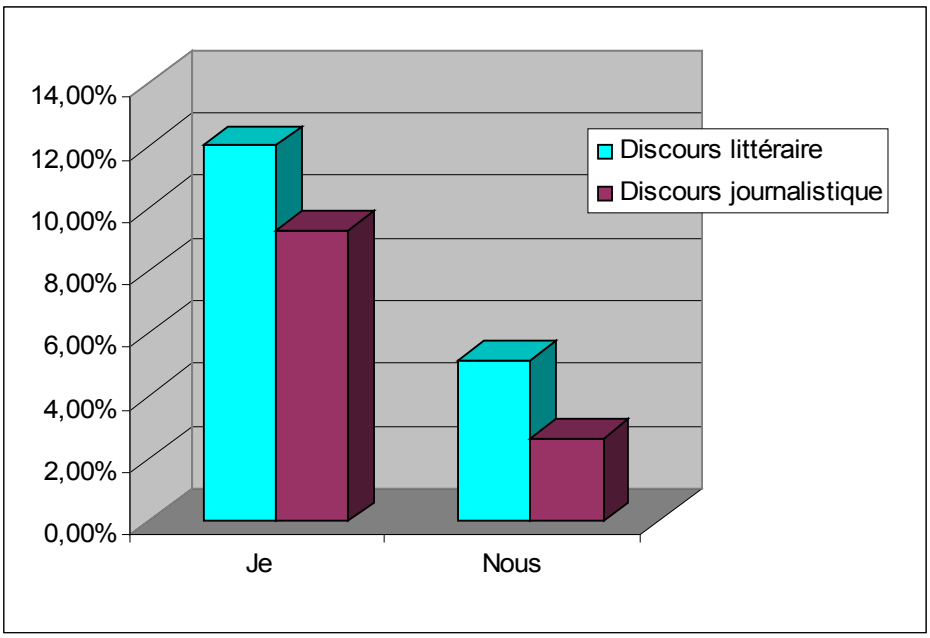


Fig. 6 : % de pronoms personnels « je » et « nous » parmi l'ensemble des pronoms personnels

Nous constatons que la grande majorité des occurrences est donnée par le discours littéraire, qui exploite de façon dominante la première et la deuxième personne. Cela est lié encore une fois à la

spécificité du discours. L'article littéraire s'inscrit dans un discours fortement conditionné par des contraintes normatives. Ce n'est pas une question de choix, comme tout énonciateur le fait lorsqu'il s'approprie le langage, mais une contrainte dictée par le nature du discours. La première personne permet ainsi d'opposer le point de vue critique-chercheur aux autres travaux précédents qui s'intéressent à la question étudiée. Et contrairement à ce qui a été montré par Loffler-Laurian[1] dans son étude sur le discours scientifique en chimie et en physique, le discours des études littéraires fait un usage systématique de la première personne car l'aspect des théories est subjectif et la valeur des hypothèses est personnelle. Or dans l'article journalistique, il n'est question ni de théories ni d'hypothèses. On écrit pour être lu par un large public. Le point de vue du journaliste importe peu, un spécialiste du cinéma n'intéressera que peu de lecteurs, les érudits comme lui, alors que dans le discours littéraire, la démarche est inverse : l'article, qui est l'image d'une école de pensée, expose avant tout la rigueur méthodologique et les préalables à la découverte scientifique. On s'adresse à des spécialistes du même domaine, on cherche à partager avec des collègues les résultats d'une recherche dans un langage spécialisé, on est loin de la vulgarisation.
Chaque discours a donc une forme canonique au niveau des pronoms, et les temps verbaux permettent de mieux comprendre l'usage qui en est fait. Les pronoms sont étroitement liés à une autre catégorie grammaticale, celle des verbes, dont la distribution dans notre corpus est représentée par le graphique suivant :

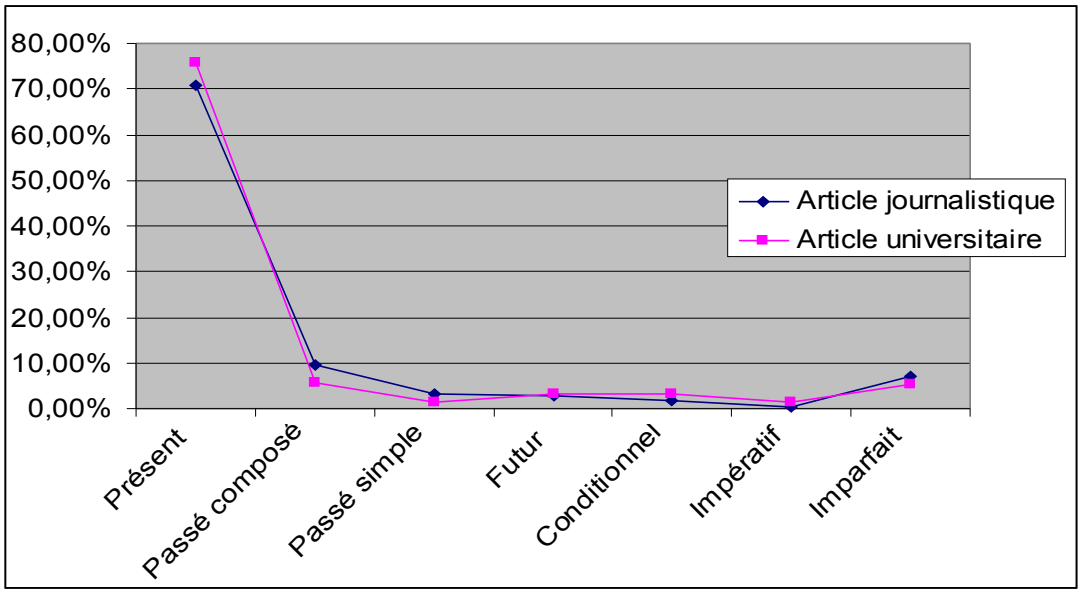


Fig. 7 : La distribution des temps verbaux dans le corpus

Avec l'aide de l'étiquetage de Cordial, nous avons pu effectuer une analyse de l'ensemble des deux discours. La distribution des différents modes de notre corpus montre un indicatif qui domine largement. Quant aux autres modes, ils sont minoritaires. Nous ne nous intéresserons ici qu'à certaines des sous-catégories qui nous ont semblées les plus discriminantes.
Au niveau des temps verbaux, nous observons premièrement une opposition au niveau de l'axe passé *vs* présent. Le discours journalistique se caractérise par un usage plus important des temps narratifs, comme pour le passé composé, (9,69% contre 5,63%), le passé simple, (3,04% contre 1,58%) et l'imparfait (6,91% contre 5,30%), tous les trois exprimant une temporalité passée. Quant à l'axe du présent, c'est le discours littéraire qui marque les plus grandes fréquences : un mode narratif à visée atemporelle typique des textes scientifiques, et d'ailleurs caractérisé par l'usage du présent de l'indicatif à hauteur de 75,87% contre 70,86% pour le discours journalistique. L'emploi du futur, dominant chez le littéraire, souvent dans l'introduction, avec une fréquence de 3,31%, est lié à des contraintes stylistiques, comme dans des expressions de ce genre : « nous nous attacherons… », « nous verrons ensuite », « nous nous limiterons », qui permettent d'anticiper sur les résultats de la recherche. Un dernier mode qui nous semble le plus discriminant est l'impératif.

[1] Cf. l'auteur 1980, p.135.

Bien que l'impératif soit un mode très limité dans notre corpus, 0,30% dans le discours journalistique contre 1,35% pour le discours littéraire, sa distribution n'est pas sans intérêt. Le littéraire écrit souvent dans le but d'être compris, et il veut surtout s'assurer que son lecteur le suive. D'où l'emploi fréquent de l'impératif à des fins pédagogiques et démonstratives : « Ajoutons qu'au niveau dialogique… », « Commençons par cette dernière question… », « regardons de plus près la succession de… », « passons aux opérations interprétatives ».

Conclusion(s)

L'objectif de cette étude n'était pas de dresser un parangon de chacun des deux discours. Elle n'avait pas non plus la prétention de traiter de tous les aspects sous lesquels on pourrait définir le genre de la critique. Mais elle aura montré que l'insertion d'un genre dans un discours donné n'est pas sans influence sur l'aspect morphosyntaxique et stylistique du texte. Car le facteur prédominant de ces divergences semble être celui du discours. En effet, le profil morphosyntaxique qui émerge de nos différentes analyses est celui d'une écriture qui exprime, en fonction de la situation des discours, à la fois les spécificités et la diversité du genre. Un genre oscillant entre une écriture concise et technique, entre un style privilégiant les mots courts et les points, la troisième personne, le passé composé et le passé simple (l'article journalistique), et une écriture pédagogique et normative (l'article littéraire), préférant les mots longs et la virgule, la première personne, le présent et l'impératif. Ces résultats nous permettent de voir le bien fondé des variables morphosyntaxiques dans la définition contrastive d'un profil-type de la critique, qui sont à même de donner une représentation objective des genres en fonction des discours dans lesquels le texte prend place. C'est donc sur une typologie des discours que se fondera une typologie des genres, à travers laquelle nous pourrons regrouper et typer les textes.

BIBLIOGRAPHIE

BEAUVISAGE, T. 2001. Exploiter des données morphosyntaxiques pour l'étude statistique des genres - Application au roman policier, *TAL,* 16.

BIBER, D., CONRAD, S., REPPEN, R. 1998. *Corpus linguistics, Investigating Language, Structure and Use*, Cambridge, Cambridge Approaches to Linguistics.

BRUNET, E. (éd.) 1983. *Etude statistique des textes littéraires, Hommage à Pierre Guiraud,* Cumfid n°14, CNRS, Institut national de la langue française, URL 9 - Université de Nice octobre 1983.

BRUNET, E. 1988. *Le vocabulaire de Victor Hugo*, Paris-Genève, Champion-Slatkine.

LOFFLER-LAURIAN, A.-M. 1980. L'expression du locuteur dans les discours scientifiques."JE", "NOUS" et "ON" dans quelques textes de chimie et de physique, *Revue de linguistique* romane, 44, pp. 135-157.

LOISEAU, S. POUDAT, C. & ABLALI, D. 2006. Exploration contrastive de trois corpus de sciences humaines, *JADT*, Besançon, Les cahiers de la MSH Ledoux, pp. 631-642.

MALRIEU, D. & RASTIER, F. 2001. Genres et variations morpho-syntaxiques, in Daille, Romary (dir.), *Linguistique de corpus*, *TAL*, vol. 42 n°2, Paris, Atala/ Hermès, pp. 547-577.

MULLER, C. 1979. *Langue française et linguistique quantitative, Recueil d'articles*, Genève, Éditions Slatkine.

RASTIER, F. 2001. Eléments de théorie des genres, in *http://www.revue-texto.net*.

RASTIER, F. 2001. *Arts et sciences du texte*, Paris, PUF.

CORPUS ET DIACHRONIE : DE LA CONSTITUTION AU TRAITEMENT
Un cas d'espèce : le roman sentimental moderne

Magali BIGEY
Université de Franche-Comté (LASELDI)

SOMMAIRE

Résumé : *La constitution d'un corpus de littérature sérielle pose différents problèmes, dont ceux du recueil de données et des droits de reproduction. Une fois ces problèmes écartés, se pose alors la question du traitement. Quel type appliquer au corpus, dans quel but, quels résultats peut-on attendre ?*
Ainsi, après une brève présentation du thème de recherche, nous évoquerons les traitements effectués (création d'AFC avec le logiciel ASTARTEX, création de dictionnaires électroniques spécifiques à l'aide du logiciel Nooj) sur ce corpus réunissant 50 romans numérisés, publiés de 1978 à 2004.

1. Roman sentimental et littérature sérielle

Ce travail porte sur l'évolution du roman sentimental depuis 1942. Le corpus est divisé en deux parties, constituées de romans sentimentaux de type sériel. La première partie du corpus est le support d'une analyse narratologique[1], la seconde d'une analyse lexicologique, le tout en diachronie.
L'objet qui nous intéresse aujourd'hui est la seconde partie du corpus, qui est la partie numérisée.
Le roman sentimental moderne appartient à la littérature sérielle. Son apparition en France remonte à 1978, avec l'arrivée du roman Harlequin, parangon du genre encore aujourd'hui.
Cette littérature voit des éditions multiples, qui ne restent sur le marché que deux ou trois semaines en moyenne, avant d'être retirées de la vente et détruites, pour être aussitôt remplacées par d'autres.
Ils sont tous (ou presque) traduits de l'anglais, mais la traduction n'est pas l'objet de ce travail. Nous avons construit notre corpus à partir des traductions françaises des romans.

[1] Le schéma narratif canonique de référence est celui dégagé par Julia Bettinotti et son équipe.

2. Le corpus

2.1. Du corpus papier…

Pour la constitution de ce corpus diachronique, la première difficulté a été de trouver des romans parfois disparus depuis plusieurs décennies.
Le recours aux petites annonces et autres foires aux livres et brocantes a permis de réunir plusieurs centaines de romans, publiés de 1978 à 2004.
Ce type de démarche est un passage obligé pour toute recherche en littérature sérielle.
A la suite de cette collecte, cinquante romans ont été retenus. Le principal critère de sélection, pour les romans datant d'avant 1990, est leur date de parution.
Le corpus réuni couvre une période de 26 ans.
Le résultat est un corpus de 8000 pages papier, très peu exploitable. C'est pour cette raison que nous avons eu recours à la numérisation.

2.2. Au corpus numérisé

Afin d'avoir un corpus complet, il a fallu numériser chaque page de chaque roman.
Un logiciel de reconnaissance automatique de caractères (OCR) a permis d'obtenir un corpus presque exploitable et analysable par des logiciels de traitement automatique de textes.
Le corpus final de 2 millions de mots représente 3146 pages Word[1].

Nettoyage du corpus

L'OCR n'a pas donné un texte exploitable immédiatement, et la phase de préparation a été encore longue. Un tel corpus doit être nettoyé de ses « coquilles » et « mots inconnus » afin de pouvoir être utilisé.

Corrections

La première étape de correction s'est faite sur écran. Chaque page numérisée est relue et débarrassée des éléments « inconnus » dus à l'océrisation, mais beaucoup d'entre eux passent à côté de la vigilance du lecteur.
Il reste encore beaucoup d'éléments non reconnus, identifiés grâce au logiciel Nooj[2]. Est alors créée une liste des mots inconnus.
Ces mots inconnus ont diverses origines et doivent être traités de manière différente.
Certains pourront être corrigés automatiquement, d'autres nécessiteront des traitements au cas par cas.

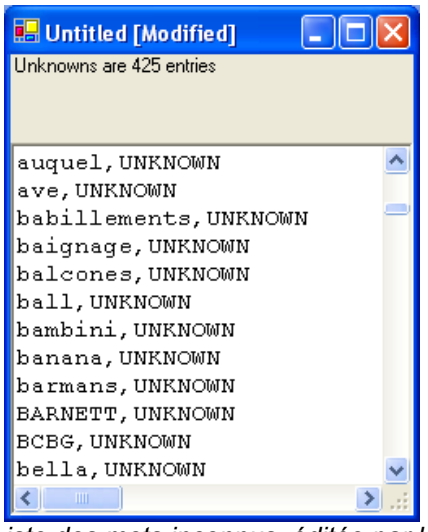


Fig. 1a : Liste des mots inconnus, éditée par le logiciel.

Les mots inconnus sont de plusieurs types :
- les mots inconnus de type *« coquilles »*, soit dus à la numérisation (le logiciel de reconnaissance automatique reconnaît souvent un *« rn »* pour un *« m »*, un *« sreur »* pour *« sœur »*, un *« 1 »* pour *« l »* ex : *panta1on*) et qui ont échappé à la correction manuelle, soit des *« coquilles »* présentes

[1] Pages format Word, police Times New Roman, 12 points.
[2] Ce logiciel de traitement automatique est développé par Max Silberztein, laboratoire LASELDI, université de Franche-Comté.

dans l'exemplaire papier du roman. (ex : *« baignage »* pour *« baignade »*, *« corgase »* pour *« corsage »*)
- problème de découpage des mots par le logiciel. On trouvera par exemple *« er »* considéré comme une erreur, alors qu'il appartient au mot *« amer »*.
Pour la correction des coquilles du type *« sreur »*, nous avons utilisé la fonction « rechercher-remplacer » de Word, qui permet une correction automatique des différentes occurrences.
- les mots inconnus qui sont en fait des mots étrangers (anglais, espagnols, tibétains…).
Ils devront être traités à part, mais il est essentiel de les garder intacts dans le texte.
- les mots inconnus qui sont des « mots d'enfants », tel *« éphélan »* pour éléphant, ou encore des mots « inventés » par le traducteur, sortes de néologismes ou d'abus de langage tels *« funambuliste »* ou *« crispement »*[1].
- les mots inconnus qui correspondent à des retranscriptions d'hésitations dans des dialogues : tels : *« abso… »* pour *« absolument »*, *« tre »* pour *« en-tre »*.
- les mots inconnus des dictionnaires électroniques mais pas du TLFi[2] tels *« babillements »* ou *« poincianas »* (qui est une plante).

Certaines de ces formes sont intégrées à un dictionnaire filtre[3] afin d'être reconnues par le logiciel. Le dictionnaire est spécifique à ce corpus. Il comprendra les mots étrangers, les noms propres non reconnus, les mots inconnus des différents dictionnaires, les abus de langage et élisions trouvées dans les dialogues. Ces mots sont importants pour l'analyse et doivent rester en l'état.
D'autres sont corrigées. Ce sont celles qui résultent de la numérisation, les coquilles dues au scanner.

3. Le traitement automatique

Parmi toutes les possibilités de traitement offertes par les logiciels de traitement automatique, il faut choisir la méthode adaptée au type de travail souhaité.
Plusieurs possibilités s'offrent à nous :

3.1. La lemmatisation

Très efficace pour les analyses de champs sémantiques, elle regroupe sous un même lemme des formes graphiques différentes. Malheureusement, ce type de traitement entraîne une perte d'information sémantique :
Exemple : *vouloir*, *veux*, *voudrais*, *aurais voulu*… ont des sens très différents, et seraient pourtant réunis sous une même « étiquette ».

3.2. L'étiquetage

Cette méthode consiste à donner à des formes graphiques une étiquette relevant de leur statut. Il existe les étiquetages morphologiques et morphosyntaxiques.
Un des avantages de l'étiquetage est qu'il permet de faire des recherches très précises du type : *« tous les verbes au passé simple »*, *« toutes les occurrences du verbe aimer au présent »* ou *« tous les noms de parties du corps »…*

3.3. L'Analyse Factorielle des Correspondances (A.F.C.)

> « L'analyse factorielle traite des tableaux de nombres et elle remplace un tableau difficile à lire par un tableau plus simple à lire qui soit une bonne approximation de celui-ci. »[4]

L'analyse factorielle donne une « cartographie » de la répartition du vocabulaire dans le corpus et isole des phénomènes qui seraient peut-être passés inaperçus.

---

[1] Ces mots « inventés » sont quasi inexistants depuis la fin des années 80. Les traducteurs sont aujourd'hui des personnes qui ont fait de hautes études littéraires et qui parlent parfaitement l'anglais.
[2] Trésor de la Langue Française informatisé, http://atilf.atilf.fr
[3] Ce dictionnaire est ensuite intégré au logiciel Nooj.
[4] Cette citation est tirée du Que sais-je ? de Philippe Cibois, « L'analyse factorielle : analyse en composantes principales et analyse des correspondances», p.5.

En effet, il est difficile de visualiser la différence entre deux extraits de textes qui paraissent identiques dans leur style et leur vocabulaire :

> Exemples :
> Une légère brise agita les voilages de mousseline. Dehors, les criquets, cachés dans l'herbe haute, firent entendre leur cri strident et monotone. La lumière dorée du couchant filtrait à travers les tentures. Cette soirée de printemps ressemblait à des milliers d'autres, et pourtant je savais qu'elle était différente. Elle marquait la fin d'une époque.[1]
> Si l'on se fiait aux brochures, l'île ne mesurait pas plus de huit kilomètres de long. Mais les routes avaient été tracées de manière à respecter au mieux la forêt, si bien qu'il fallut zigzaguer pendant un quart d'heure avant d'atteindre le parking de l'auberge Seagrass. Un portier en livrée bâillait copieusement sur le seuil.[2]

Les extraits ci-dessus sont issus de deux textes, numérotés 17 et 23 sur la représentation suivante.

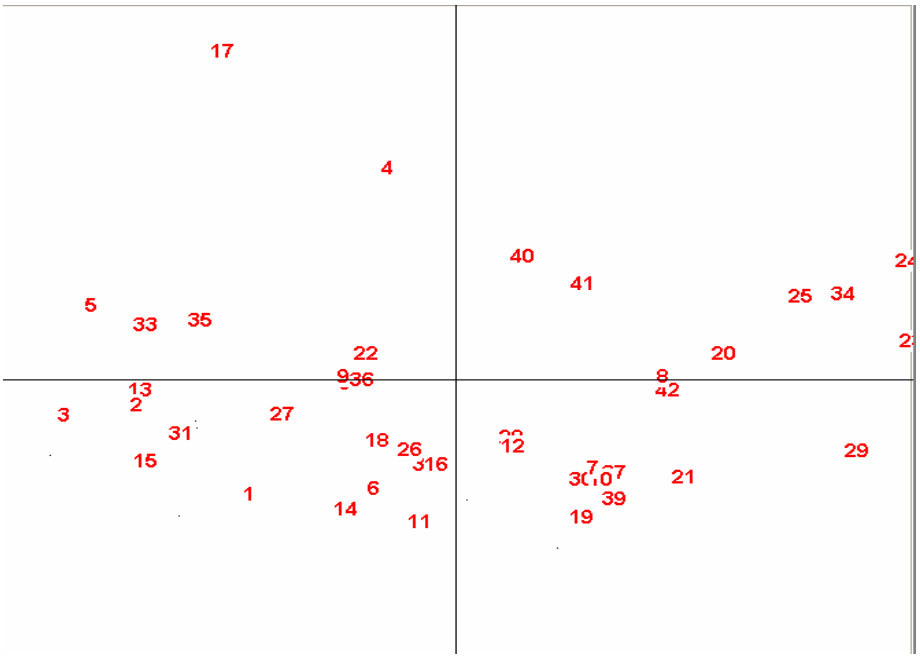


Fig. 1b : AFC de la répartition des romans en fonction de leur vocabulaire[3]

Un simple regard sur le résultat de l'AFC donne des directions de recherche. Ici, plusieurs hypothèses s'offrent à nous.
Après vérification, il s'est avéré que le texte 17 est le seul du corpus à avoir une orientation policière, et on peut supposer que c'est ce qui explique sa situation dans le tableau d'analyse. Mais en y regardant de plus près, on peut voir que ce texte 17 a un narrateur homodiégétique, ce qui est très rare dans le roman sentimental sériel. Ces romans sont très souvent constitués essentiellement de dialogues et le narrateur est la plupart du temps hétérodiégétique, ce qui est le cas pour le texte 23.
Une autre indication est donnée par le groupe formé à droite de l'AFC. Il s'avère après vérification que les romans 20, 24, 25 et 29 ont tous été écrits par le même auteur, Penny Jordan.
Cette indication oriente une nouvelle piste de recherche.

3.4. Le choix final

Nous avons finalement choisi de travailler sur un texte non lemmatisé, pour limiter la perte d'information sémantique, mais ce choix augmente considérablement le nombre des hapax dans la liste de vocabulaire.

---

[1] Rebecca Flanders, 1994, *« Vertigo »*, Sixième Sens, Harlequin, Paris, p.6.
[2] Regan Forest, 1990, *« La maison du cauchemar »*, Suspense, Harlequin, Paris, p.1.
[3] Cette AFC est issue du logiciel ASTARTEX, développé par Jean-Marie Viprey, laboratoire LASELDI, université de Franche-Comté.

Nous travaillons aussi sur une version du texte étiqueté morphologiquement, de manière automatique (par le logiciel Nooj). Nous avons décidé d'ignorer pour l'instant la marge d' « erreur » d'étiquetage, pour une partie du travail qui consiste en une recherche des variations sociologiques et linguistiques du vocabulaire en diachronie.

4. L'analyse par dictionnaires électroniques spécifiques

La création de dictionnaires électroniques spécifiques permet d'adapter l'outil au corpus à traiter. Cela permet aussi des recherches d'occurrences et de co-occurences plus précises, car les éléments sont codés individuellement.

4.1. Le vocabulaire des parties du corps

La fouille de ce corpus a fait émerger des tendances d'utilisation de vocabulaire. Après quelques temps, il a paru évident que le vocabulaire des parties du corps devait faire l'objet d'une étude à part entière.

4.2. Préparation du dictionnaire

Les dictionnaires sont créés pour être intégrés au logiciel Nooj.
La liste du lexique des parties du corps a été réalisée à partir de la liste totale du lexique du corpus.

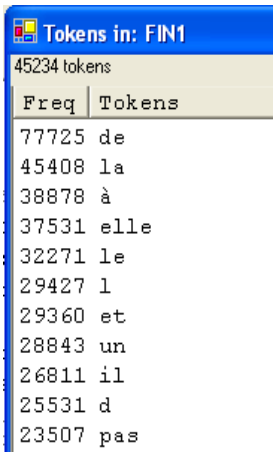


Fig.2 : Extrait de la liste des formes du corpus

Il est possible de classer le lexique par ordre alphabétique ou par nombre d'occurrences.
Nous pouvons voir ici que le texte comporte 45234 formes différentes.
Nous avons passé en revue environ 36000 occurrences (nous avons laissé de côté les hapax), afin de ne retenir que les termes utilisés pour désigner les parties du corps. Nous avons aussi laissé de côté les usages métaphoriques de certains termes, qui pourraient faire l'objet d'un autre travail.
Au final, une liste de 221 entrées a été faite. Elle constitue la base du dictionnaire électronique.
Nous tenons à signaler ici l'importance de travailler sur un texte non lemmatisé. Nous traitons par deux entrées, par exemple *« rein »* et *« reins »*.
Il est important de les différencier, car dans ce cas, l'utilisation de *« rein »* est réservée au domaine de la maladie (greffe de rein) et de la douleur[1], alors que l'utilisation de *« reins »* est réservée à la description du corps avec *« chute de reins »*, *« creux des reins »*… Dans ce deuxième cas, seules deux occurrences sur 46 font référence au domaine médical.

4.3. Création du dictionnaire

Une fois la liste du lexique des parties du corps établie, il faut coder le dictionnaire. Chaque occurrence doit pourvoir être reconnue comme *« Partie du corps »* par le logiciel.
Les codes sont propres à chaque dictionnaire et il suffit d'inventer son propre code en fonction de ce qu'on souhaite rechercher.

[1] Dans le corpus, sur 36 occurrences de *« rein »*, une seule n'est pas rattachée au domaine médical.

Le premier code à appliquer est le code « Parties du corps », afin que chaque terme soit reconnu comme tel. Le code choisi est PdC.
Ensuite, il faut ajouter d'autres codes.
Ici, les choix suivants ont été faits :
+N : nom
+f : féminin
+m : masculin
+s : singulier
+p : pluriel
+sup : partie supérieure du corps
+inf : partie inférieure du corps
+int : à l'intérieur du corps
+mem : partie des membres
+vis : partie du visage
+tet : partie de la tête
+y : yeux
+sex : sexuel
Il est très important de surcoder un dictionnaire, afin de pourvoir faire des concordances sur des demandes bien spécifiques.
Si le dictionnaire codé ne présentait que « PdC », avec genre et nombre, comme on pourrait s'y attendre, il serait impossible de faire une recherche sur tous les mots du lexique qui décrivent la tête, ou le bas du corps, ou encore le champ sémantique descriptif des yeux…

Extrait du dictionnaire :
annulaire,N+m+s+sup+mem+PdC
articulations,N+f+p+int+PdC (…)
chevelure,N+f+s+sup+tet+PdC
cheveu,N+m+s+sup+tet+PdC
cheveux,N+m+p+sup+tet+PdC (…)
nombril,N+m+s+inf+sex+PdC (…)
nuque,N+f+s+sup+PdC (…)
œil,N+m+s+sup+tet+vis+y+PdC

Une fois la liste établie dans un fichier Word, il suffit de la copier et de l'insérer dans un fichier spécifique. Après compilation, le dictionnaire est utilisable.

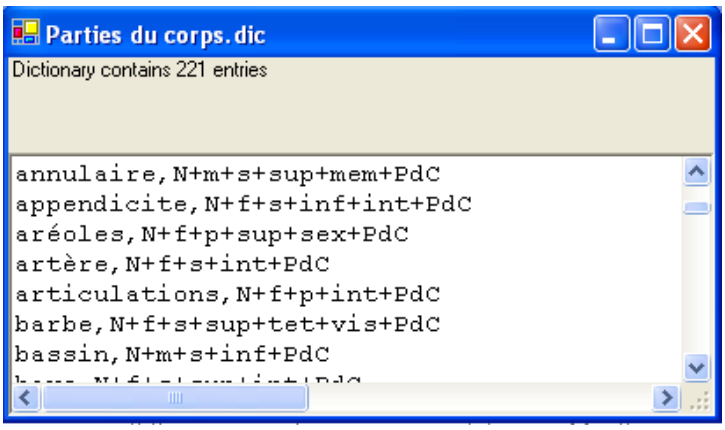


Fig. 3 : Extrait du dictionnaire

Par exemple, nous souhaitons travailler sur tous les noms féminins qui décrivent le haut du corps :
Il suffit d'effectuer une recherche sur les codes N,PdC, f et sup.

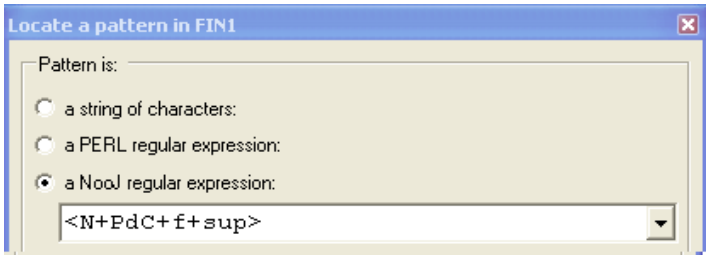


Fig. 4 : fenêtre de recherche des codes

Résultat : une liste de toutes les occurrences, en contexte, des noms féminins qui se situent dans la partie supérieure du corps.

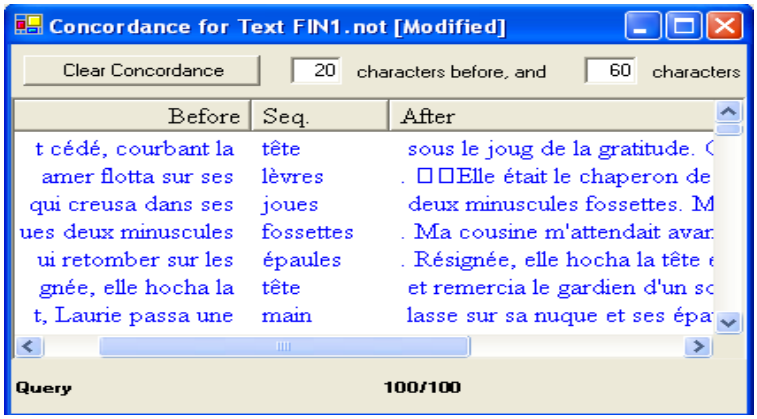


Fig. 5 : résultat de la recherche

Nous tenons à signaler que la recherche d'occurrences respecte l'ordre du texte, et présente donc les résultats en diachronie.

5. Variations en diachronie

Un corpus diachronique permet de repérer des variations, qui peuvent être de plusieurs ordres : linguistiques, sociologiques ou autres.
Un petit module statistique permet de visualiser sous forme de graphique les fréquences de chaque mot recherché.

5.1. Variations sociologiques

Exemple 1 :

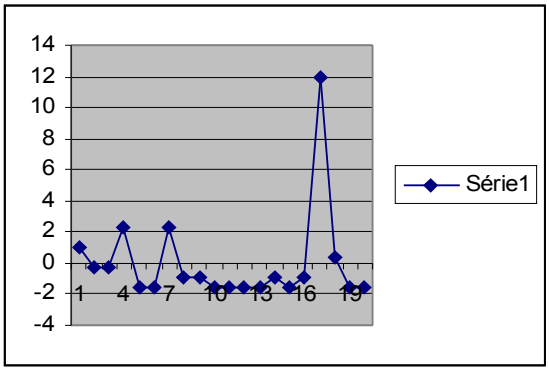


Fig. 6 : courbe représentative de l'évolution de « cigarette »

Dans cet exemple, nous remarquons la baisse d'utilisation du mot *« cigarette »* depuis les années 90, avec un seul pic important. Ce pic correspond à un roman dont le héros est dépressif, et fume beaucoup.

Exemple 2 :

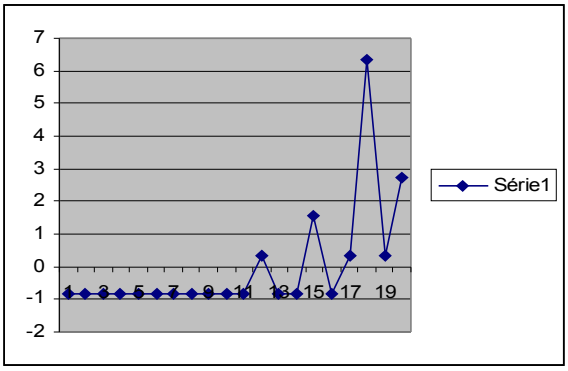


Fig. 7 : courbe représentative de l'évolution du syntagme « ordinateur portable »

Le roman sentimental appartient au genre populaire, et suit de près l'évolution de la société.
Ce phénomène explique l'importance croissante que prend le syntagme *« ordinateur portable »* à partir de la fin des années 90.

5.2. Variations linguistiques

Exemple :

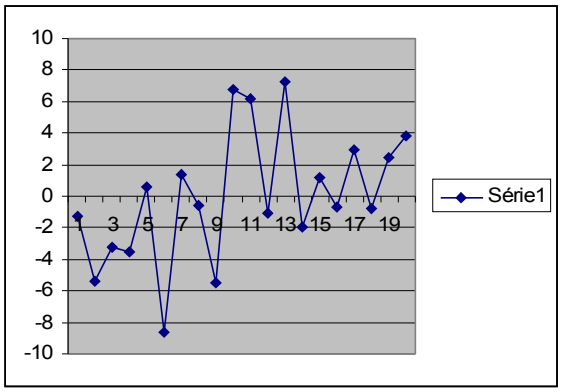


Fig. 8 : courbe représentative de l'évolution des verbes au participe passé

Une tendance d'utilisation très peu régulière se dessine. Il faut alors rechercher les occurrences en contexte afin d'en identifier la ou les causes.

Conclusion

Nous avons essayé de montrer que le choix d'un type de traitement automatique n'était pas anodin, ni forcément évident.
Parmi les différentes méthodes proposées, il en est certaines qui conviennent mieux que d'autres à l'analyse diachronique. Ici, nous avons privilégié l'étiquetage au détriment de la lemmatisation, car garder la charge sémantique des mots nous est apparu comme primordial.
Les choix de traitements vont permettre au chercheur d'orienter son travail. En quelques secondes, il lui est possible de vérifier une hypothèse, en créant par exemple une courbe représentative de l'évolution d'une forme ou d'un syntagme. Il peut alors visualiser immédiatement si l'entreprise d'une recherche est judicieuse ou non.
Le travail sur un tel corpus nécessite un traitement pluridisciplinaire. Ici, sont convoqués la linguistique-informatique, la littérature, la sociologie et le Traitement Automatique des Langues (TAL). L'important est de pouvoir faire cohabiter ces disciplines, et le traitement automatique le permet en créant un lien qui les rapproche de l'objet d'étude et dans l'objet d'étude.

BIBLIOGRAPHIE

BERNARD, M. 1999. *Introduction aux études littéraires assistées par ordinateur*, Paris, Presses Universitaires de France.

BETTINOTTI, J., (sous la dir. de) 1986, *La corrida de l'amour : le roman Harlequin*, éd. *Les cahiers du département d'études littéraires*, n°6, Montréal.

BRUNET, E. 1981. *Le vocabulaire français de 1789 à nos jours*, Genève Paris, Slatkine-Champion.

CIBOIS, P. 1983. *L'analyse factorielle : analyse en composantes principales et analyse des correspondances*, Paris, Presses Universitaires de France, Que sais-je ?.

COUEGNAS, D. 1992. *Introduction à la paralittérature*, Paris, Seuil.

HABERT, B., NAZARENKO, A. et SALEM, A. 1997. *Les linguistiques de corpus*, Paris, Armand Colin, Masson.

PEQUIGNOT, B.1991. *La relation amoureuse, analyse sociologique du roman sentimental moderne*, Paris, L'Harmattan.

SILBERSTEIN, M. 1993a. *Dictionnaires électroniques et analyses automatiques de textes : le système INTEX*, Paris, Masson.

RASTIER, F. 2001. *Arts et Sciences du texte*, Paris, Presses Universitaires de France.

VIPREY, J.-M. 2002. *Analyses textuelles et hypertextuelles des Fleurs du mal*, Paris, Champion.

GROUPEMENTS DE TEXTES ET CORPUS : POINT DE VUE DE LINGUISTE

Carine DUTEIL-MOUGEL
ATILF, Nancy-Université, CNRS

SOMMAIRE

Résumé : *Trois situations de lecture des textes littéraires sont proposées aux élèves dans l'enseignement secondaire : au sein d'anthologies littéraires, dans le cadre de groupements de textes et dans le cadre d'une œuvre intégrale.*
Les groupements de textes rassemblent des extraits (scènes de pièces de théâtre, passages de romans ou de nouvelles, extraits d'essais, extraits de poèmes ; textes tirés d'un recueil : fables, lettres, nouvelles, poèmes) choisis et étudiés selon une cohérence thématique ou problématique. Leur étude prend place dans le cadre d'une séquence didactique, et doit permettre d'aborder un objet d'étude *au programme.*
L'une des épreuves écrites du Capes interne (depuis 2001) consiste justement à analyser un groupement de textes, accompagné très souvent d'un document iconographique, et à proposer une exploitation didactique de ce groupement sous la forme d'un projet de séquence. Notre réflexion portera sur ces « groupements », appelés de plus en plus souvent « corpus ». Nous nous interrogerons sur leur statut : s'agit-il de véritables CORPUS ? Quels critères président au choix de leur constitution ? Comment leur pertinence est-elle établie ?
L'étude d'exemples attestés (groupements au sein de séquences didactiques, groupements proposés lors de l'épreuve anticipée du baccalauréat de français, groupements proposés lors de l'« épreuve de didactique » au Capes interne de Lettres Modernes) nous conduira à analyser les relations qui s'établissent entre les éléments ainsi rassemblés au sein des groupements (liens thématiques, rapports d'interprétance, …).
On s'interrogera sur les effets de la décontextualisation des extraits (incidence sur la textualité, sur les conditions d'interprétation), et on réfléchira aux moyens nécessaires à la constitution de corpus pleinement utilisables pour la caractérisation des textes.
Cette réflexion sur les « corpus » dans l'enseignement du français nous conduira à examiner plus en détail cette notion dans le cadre de la linguistique, et à présenter nos perspectives de recherche dans ce domaine (profilage et typologie des textes).

1. Un groupement de textes *nommé corpus*…

Trois situations de lecture des textes littéraires sont proposées aux élèves dans l'enseignement secondaire : au sein d'anthologies littéraires[1], dans le cadre de groupements de textes et dans le cadre d'une œuvre intégrale[2].

La notion de « groupement de textes » est récente dans l'enseignement du français ; elle apparaît pour la première fois dans le *Bulletin Officiel* du 7 juillet 1983. Son adoption est alors un moyen de renouveler les modalités de l'épreuve anticipée de français au baccalauréat. Comme le précise Michel Descotes (formateur IUFM) : « il s'agissait d'éviter l'évaluation des capacités de l'élève à

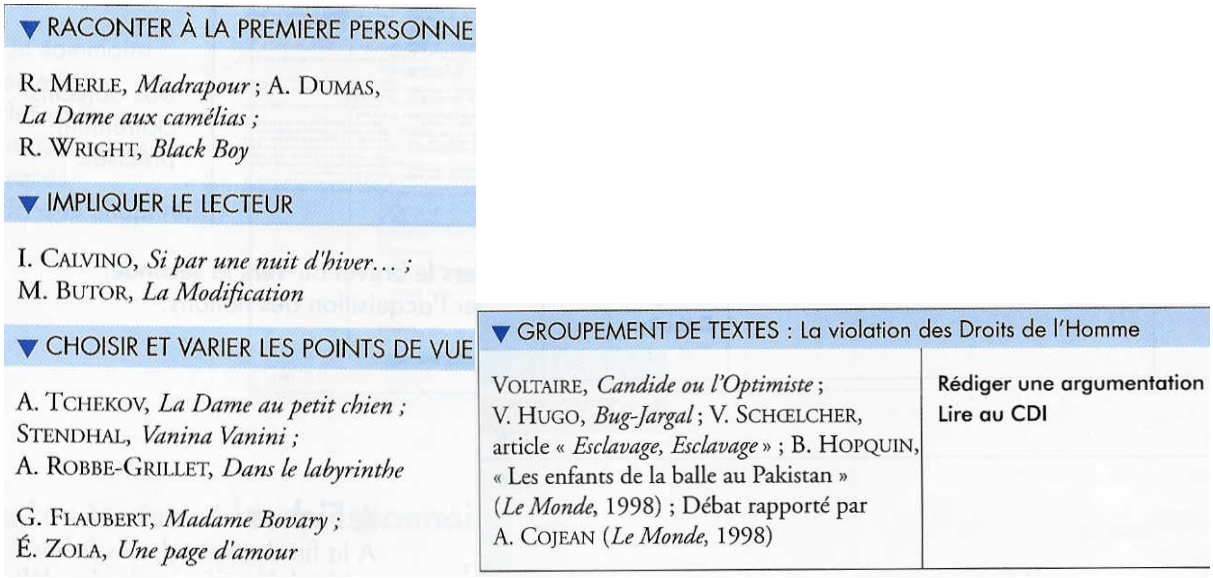
[1] Une anthologie présente des extraits d'œuvres "significatives" (textes canoniques, extraits exemplaires…). Elle s'apparente à l'encyclopédie, définie comme *archive de passages de textes décontextualisés* (cf. Rastier F., 2001, « Sémiotique et sciences de la culture », *Linx*, n°44-45, pp. 149-168).

[2] On entend par œuvre intégrale un roman, un essai, des mémoires, un recueil de poèmes, un recueil de nouvelles, une pièce de théâtre, etc. Il s'agit donc d'un texte intégral - et non d'une œuvre complète (*i.e.* l'ensemble des écrits d'un même auteur).
On soulignera que les élèves étudient beaucoup plus de groupements de textes que d'œuvres intégrales.

partir de la lecture d'un texte “à l'état de fragment isolé”. Pour cela, il est préconisé de constituer la liste d'oral avec : - des oeuvres qui ont été lues dans leur intégralité ; - des groupements de textes choisis et étudiés selon une cohérence thématique ou problématique clairement formulée : par exemple, un groupe de poèmes permettant d'étudier la fuite du temps chez les poètes romantiques ou l'automne dans la poésie d'Apollinaire ; un choix de pages groupées pour une étude de la condition humaine selon Pascal et Voltaire ou de la critique de la société chez les philosophes du XVIIIème siècle, etc. ».
À l'heure actuelle la pratique du groupement s'est développée, et le groupement apparaît de plus en plus clairement comme *un instrument didactique* permettant de construire des savoirs et des savoir-faire.

> « Réunissant un nombre limité mais suffisant de textes (quatre à cinq paraît une bonne moyenne), le groupement de textes constitue une unité d'apprentissage autour d'une notion littéraire, culturelle ou fonctionnelle. »
> *Programmes et accompagnement* (enseigner au collège, Français), 2004 (rééd.), Paris, Centre National de Documentation Pédagogique.

Le groupement de textes rassemble des extraits (scènes de pièces de théâtre, passages de romans ou de nouvelles, extraits d'essais, extraits de poèmes ; textes tirés d'un recueil : fables, lettres, nouvelles, poèmes) choisis et étudiés selon une cohérence thématique ou problématique. Son étude prend place dans le cadre d'une séquence didactique[1], et doit permettre d'aborder un *objet d'étude* au programme.
Par exemple, au collège, les groupements de textes permettent d'étudier les différentes *formes de discours* au programme[2] – l'objectif majeur des nouveaux programmes pour l'enseignement du français au collège étant la maîtrise des discours[3].



▼ RACONTER À LA PREMIÈRE PERSONNE

R. MERLE, *Madrapour* ; A. DUMAS, *La Dame aux camélias ;* R. WRIGHT, *Black Boy*

▼ IMPLIQUER LE LECTEUR

I. CALVINO, *Si par une nuit d'hiver… ;* M. BUTOR, *La Modification*

▼ CHOISIR ET VARIER LES POINTS DE VUE

A. TCHEKOV, *La Dame au petit chien ;* STENDHAL, *Vanina Vanini ;* A. ROBBE-GRILLET, *Dans le labyrinthe*

G. FLAUBERT, *Madame Bovary ;* É. ZOLA, *Une page d'amour*

▼ GROUPEMENT DE TEXTES : La violation des Droits de l'Homme

VOLTAIRE, *Candide ou l'Optimiste* ; V. HUGO, *Bug-Jargal* ; V. SCHŒLCHER, article « *Esclavage, Esclavage* » ; B. HOPQUIN, « Les enfants de la balle au Pakistan » (*Le Monde*, 1998) ; Débat rapporté par A. COJEAN (*Le Monde*, 1998)

Rédiger une argumentation
Lire au CDI

Fig. 1 : Extraits d'un manuel de Français pour la classe de troisième, Hatier, 1999, pp. 6-7
à Gauche : « Les formes de discours, 1. Raconter, décrire »
à Droite : « Les formes de discours, 2. Argumenter convaincre »

Notons que lesdites *formes de discours* renvoient aux *séquences textuelles* définies par la linguistique textuelle[4].

---

[1] Les nouveaux programmes favorisent le travail par séquences didactiques. On désigne par là « un mode d'organisation des activités qui rassemble des contenus d'ordre différent autour d'un même objectif, sur un ensemble de plusieurs séances ».

[2] « **Choix théorique : le discours**
Le premier objectif de l'enseignement du français est de “donner aux élèves la maîtrise des principales formes de discours“ (*BO* n° 10, 15 octobre 1998, p. 4). », *Programmes et accompagnement* (enseigner au collège, Français), 2004.

[3] « Le français se fixe comme objectif central au collège la maîtrise des discours. », *ibid.*

[4] Cf. notamment Adam J.-M., 1992 et 1999.

« **Formes de discours.** Suivant les finalités de l'énonciation, les discours adoptent des dominantes différentes : c'est en ce sens qu'on parle ici de discours narratif, descriptif, explicatif, argumentatif, pour les formes les plus évidentes.
– Le discours narratif rapporte un ou des événements et les situe dans le temps.
– Le discours descriptif vise à nommer, caractériser, qualifier.
– Le discours explicatif cherche à faire comprendre.
– Le discours argumentatif valorise un ou plusieurs points de vue, une ou plusieurs thèses.
On peut trouver diverses formes de discours dans un même texte. »
Glossaire, *Programmes et accompagnement* (enseigner au collège, Français), 2004 (rééd.), Paris, Centre National de Documentation Pédagogique

La progression d'ensemble[1], de la 6ème à la 3ème, s'organise ainsi autour de deux pôles : le pôle narratif et le pôle argumentatif, que nous représentons comme suit :

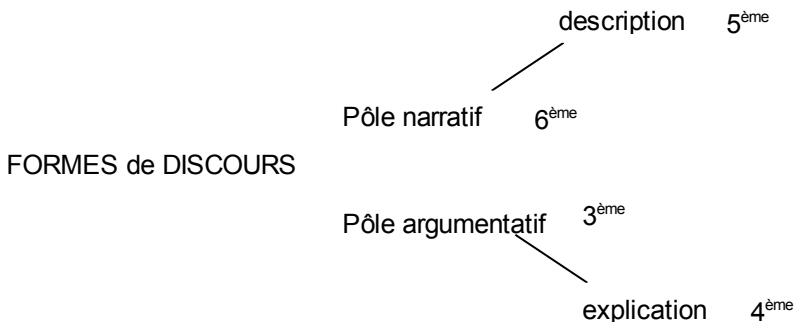


Fig. 2 : Etude des formes de discours au collège

L'appellation « corpus » pour de tels groupements apparaît à plusieurs reprises dans les nouveaux programmes pour le secondaire. Cette appellation figure également dans les modalités et les libellés des sujets de l'épreuve anticipée (écrite) de français au baccalauréat :

« Épreuve écrite
Les sujets prennent appui sur un ensemble de textes (corpus) distribués au candidat, éventuellement accompagnés par un document iconographique si celui-ci contribue à la compréhension ou enrichit la signification de l'ensemble. »
Note de service n° 2001-117 du 20 juin 2001. BO n° 26 du 28 juin 2001
Modifiée par la note de service n° 2003-002 du 8 janvier 2003. BO n° 3 du 16 janvier 2003

Les sujets invitent les candidats à s'interroger sur l'unité du corpus et à réfléchir au rapprochement des extraits choisis (cf. l'exemple ci-dessous). Il peut s'agir d'un thème commun, d'un genre commun, d'un registre commun, d'un dispositif énonciatif commun, de l'appartenance à un même mouvement littéraire, etc.

***SUJET : Série ES, S – Juin 2003**
Objet d'étude : La poésie
CORPUS
Texte A. Paul Verlaine, « Mon rêve familier », *Poèmes saturniens*, 1866.
Texte B. Robert Desnos, « J'ai tant rêvé de toi », in « À la mystérieuse », *Corps et Biens*, 1930.
Texte C. Paul Eluard, « La dame de carreau », *Les Dessous d'une vie*, 1926.
Texte D. Claude Roy, « Tant », *Le Voyage d'automne*, 1987.

I. Question
Justifiez le rapprochement des quatre poèmes.
II. Ecriture : au choix
1. Commentaire : Vous ferez le commentaire du poème de Robert Desnos (texte B.)

[1] Les formes de discours sont progressivement plus complexes (dominantes, insertions, textes mêlant narratif et argumentatif, etc.). L'étude du récit est privilégiée en 6ème. En 5ème l'accent est mis sur la description ; en 4ème, sur l'explication. Les principales formes d'argumentation sont étudiées en 3ème.

2. Dissertation : Pour Eluard, le poète « aimant l'amour » n'est pas tant amoureux d'une femme que de l'amour lui-même.
La vocation de la poésie est-elle, selon-vous, de célébrer l'amour ou privilégiez-vous d'autres fonctions ?
Vous vous appuierez pour répondre à cette question, sur les textes du corpus et les poèmes que vous avez lus et étudiés.
3. Invention : Dans la préface d'une anthologie de poèmes d'amour que vous avez réunis, vous démontrez comment l'inspiration poétique et l'amour sont à vos yeux liés.
Rédigez la préface.
Vous devez nourrir votre texte de citations de poèmes et de références à des auteurs.

La nouvelle épreuve écrite du Capes interne de Lettres Modernes (depuis 2001) – l'« épreuve de didactique » – consiste également à analyser un corpus, accompagné d'un **document** iconographique (cf. les deux sujets ci-dessous), et à proposer une exploitation didactique de ce corpus sous la forme d'un projet de séquence.

« Epreuve de didactique de la discipline. Un corpus de textes éventuellement accompagné de documents iconographiques est proposé aux candidats.
Ceux-ci, dans un devoir rédigé et argumenté :
- analysent les textes, en fonction d'une problématique indiquée par le sujet ;
- proposent une exploitation didactique de ces textes, sous la forme d'un projet de séquence destinée à la classe de collège ou de lycée indiquée par le sujet. Il appartient au candidat de déterminer l'objectif qu'il fixe à sa séquence. Une séance d'étude de la langue est obligatoirement comprise dans cette séquence.
Durée de l'épreuve : Six heures ; coefficient : 1.
Le programme des épreuves est celui des lycées d'enseignement général et technologique et des collèges. »
Extrait de l'Arrêté du 2 mars 2000

***Exemple 1 : session 2001**
SUJET
Vous analyserez les documents joints dans la perspective de l'étude du récit autobiographique dans une classe de 3e, en vous attachant à la diversité des approches et des projets de chaque auteur. Vous en proposerez une exploitation didactique sous forme d'un projet de séquence, incluant une séance d'étude de la langue et des prolongements par des lectures cursives.

Documents extraits de[1] **:**
- Chateaubriand, *Mémoires d'Outre-Tombe* [Livre premier, de « Il y a quatre ans qu'à mon retour de terre sainte, j'achetai près du hameau d'Aulnay... » à « ... Et enfin l'Histoire des grands officiers de la couronne du P. Anselme. »]
- M. Yourcenar, *Archives du Nord* [éd. Gallimard, 1977, de « Dans un volume destiné à former avec celui-ci les deux panneaux d'un diptyque... » à « Survolons-le à une époque où il était encore sans habitants et sans nom. »]
- P. Néruda, *J'avoue que j'ai vécu* [éd. Gallimard, traduction de Claude Couffon, 1975, de « Je dirai pour commencer cette évocation des jours et des années de mon enfance que le seul personnage... » à « La vie était dure pour les petits agriculteurs du centre du pays. »]
- A. Nothomb, *Métaphysique des tubes* [éd. Albin Michel, 2000, de « En me donnant une identité, le chocolat blanc... » à « On commença à m'appeler par mon prénom. »]
- R. Queneau, *Chêne et chien* [1937, De « Je naquis au Havre un vingt et un février» à «je n'eus jamais de sœur. »]

Iconographie : P. Gauguin, Autoportrait au chris

***Exemple 2 : session 2006**
SUJET
Dans le cadre de l'étude de la poésie en classe de troisième, vous analyserez le corpus ci-joint en vous attachant à la diversité des démarches et des écritures.

[1] Notons qu'en 2001, il n'était pas encore question de CORPUS.

En vous fondant sur cette analyse, vous proposerez une exploitation didactique clairement problématisée tenant compte des objectifs d'étude fixés pour la classe de troisième. Votre séquence comportera obligatoirement une séance d'étude de la langue.
Vous pouvez enrichir votre projet de textes ou de références complémentaires.

CORPUS :
Victor Hugo, "Fonction du poète", *Les rayons et les ombres* (extrait), 1840.
Tristan Corbière, "Le crapaud", *Les Amours jaunes*, 1873.
Léon-Paul Fargue, *Haute solitude*, extrait d'"Azazel", 1941.
Max Jacob, *Derniers poèmes en vers et en prose*, "Amour du prochain", 1944.
Yves Bonnefoy, *Ce qui fut sans lumière*, "La tâche d'espérance", 1987.
Philippe Jaccottet, *Leçons*, "Autrefois...", in *Poésies 1946-1967*.
Jacques Réda, *Amen*, "Espère et tremble", 1968.

Pablo Picasso, *Femme en pleurs*, 1937.
[Reproduction tirée de Carsten-Peter Warncke, *Picasso*, Taschen, 1988]

1.1. Textualité, intertextualité et architextualité

Mais ces groupements de textes peuvent-ils prétendre au statut de corpus ?
La question du statut des groupements se pose dans la mesure où les critères qui président au choix de leur constitution ne sont pas explicités, et dans la mesure où la pertinence du regroupement des textes n'est pas véritablement discutée. Tout regroupement de textes ne mérite pas le nom de corpus, et on pourrait s'interroger notamment sur l'hétérogénéité des groupements - auteurs, œuvres, époques…
De plus, les textes s'apparentent davantage à des *morceaux choisis* (extraits, éléments d'un recueil, etc.). Se pose alors le problème de leur décontextualisation (texte - *le contexte, c'est tout le texte* - et corpus d'origine - *tout texte est situé dans son intertexte*) et de leur recontextualisation au sein du groupement.
Lorsqu'un texte change de corpus, il change inévitablement de sens, selon le principe d'architextualité : *tout texte placé dans un corpus en reçoit des déterminations sémantiques et modifie potentiellement le sens de chacun des textes qui le composent*[1]. L'interprétation[2] étant conditionnée par le *corpus*, l'intertexte nouveau crée ainsi des relations d'interprétance mutuelle et invite à de nouveaux parcours interprétatifs (liens thématiques, isotopies, corrélations) qu'il s'agit de pouvoir légitimer.

1.2. Parcours plurisémiotiques

L'intégration du document iconographique au corpus pose également des difficultés ; son exploitation en lien avec le corpus textuel relève parfois de l'artefact. Les programmes scolaires mettent pourtant l'accent sur l'étude de l'image[3], qui s'inscrit dans le cadre de l'apprentissage des formes de discours[4].

> « Fixe ou mobile, l'image est envisagée comme discours : elle raconte et décrit, mais elle a aussi un rôle explicatif ou argumentatif. »
> *Programmes et accompagnement* (enseigner au collège, Français), 2004 (rééd.), Paris, Centre National de Documentation Pédagogique.

Il s'agit par exemple de conduire les élèves :

---

[1] Cf. Rastier F. 2001, p. 92.
[2] L'interprétation est toutefois limitée puisqu'il s'agit de privilégier la construction d'une séquence didactique (en fonction du niveau de classe, et en suivant les indications et l'orientation du sujet) ; et que le grand nombre d'extraits condamne au butinage textuel.
[3] « Les enseignants de français, bien que non professionnels de l'image, sont amenés à intégrer dans leur enseignement la dimension visuelle, qui imprègne de plus en plus profondément la formation culturelle et les pratiques quotidiennes de leurs élèves. Afin de les conduire progressivement à une approche raisonnée de l'image, les professeurs travaillent en particulier sur les relations entre le langage verbal et le langage visuel, dans la perspective du discours. » *Programmes et accompagnement* (enseigner au collège, Français), 2004.
[4] « l'étude de l'image peut très facilement se développer à l'intérieur d'un groupement de textes et de documents, en particulier si l'on envisage les différentes formes d'argumentation rendues possibles par les langages non verbaux. » *ibid.*

> « à reconnaître, à identifier et à analyser les différentes fonctions discursives de l'image : esthétique, informative, descriptive, narrative, explicative, persuasive, critique et plus largement argumentative. » (*ibid*.)

Les manuels s'efforcent ainsi d'associer une image aux textes rassemblés. Mais la mise en relation texte(s)-image et la création de parcours interprétatifs plurisémiotiques est parfois problématique, comme l'illustre l'extrait de manuel qui figure en annexe.
L'image revêt alors très souvent une fonction illustrative, ce à quoi la réduit d'ailleurs le descriptif des modalités de l'épreuve anticipée (écrite) de français au baccalauréat :

> « Le document iconographique, s'il est joint au corpus, ne peut servir que de support. En aucun cas il ne sera demandé d'en faire une étude pour lui-même »
> *Note de service n° 2001-117 du 20 juin 2001. BO n° 26 du 28 juin 2001*
> *Modifiée par la note de service n° 2003-002 du 8 janvier 2003. BO n° 3 du 16 janvier 2003*

1.3. L'inspiration linguistique

Les programmes du français dans le secondaire sont sur de nombreux points sous-tendus par des apports de la linguistique[1] et il n'est pas exclu – cela est même fort probable – que l'utilisation récente du mot « corpus » dans les programmes soit liée au regain d'intérêt des linguistes pour cette notion[2] et à l'émergence depuis les dix dernières années de la *linguistique de corpus*.

Après nous être intéressée aux « corpus » de l'enseignement du français, nous examinerons plus en détail cette notion dans le cadre de la linguistique et nous présenterons nos perspectives de recherche dans ce domaine (profilage et typologie des textes).

2. Corpus et textualité

2.1. Corpus

La notion de *corpus* reçoit en linguistique différentes acceptions. Le corpus peut être défini comme *une collection de données langagières* ou comme *un échantillon de langage* ; il peut également être conçu comme *un ensemble de mots*, ou comme *un ensemble d'énoncés*, ou encore, dans notre perspective, comme *un ensemble structuré de textes*.
Mais tout ensemble de textes n'est pas un corpus. Pour B. Bommier-Pincemin (1999 : 416), le corpus doit vérifier trois types de conditions :

> « • *Des conditions de signifiance* : Un corpus est constitué en vue d'une étude déterminée (*pertinence*), portant sur un objet particulier, une réalité telle qu'elle est perçue sous un certain angle de vue (et non sur plusieurs thèmes ou facettes indépendants, simultanément) (*cohérence*).
>
> • *Des conditions d'acceptabilité* : Le corpus doit apporter une représentation fidèle (*représentativité*), sans être parasité par des contraintes externes (*régularité*). Il doit avoir une

[1] Pour ne citer que quelques exemples :
« Un futur professeur ne saurait ignorer que l'ensemble des programmes et instructions officielles s'inscrit dans un cadre théorique de référence, très précisément défini et illustré par les documents d'accompagnement des programmes du collège et du lycée. », *Rapport officiel du jury Capes interne et Caer de Lettres Modernes*, 2005 ;
« Le programme fixe comme objectif de fin de 3e la maîtrise des principales formes du discours.
On entend par **discours** toutes les mises en pratique du langage à l'oral et à l'écrit (voir *Glossaire*). La conception générale du programme se fonde donc sur l'énonciation, définie comme l'actualisation de la langue dans des situations concrètes d'utilisation. peu à peu à l'analyse des discours à dominante argumentative. » *Programmes et accompagnement* (enseigner au collège, Français), 2004 ; « Au collège, l'étude de la langue n'est pas une fin en soi, mais elle est subordonnée à l'objectif de la maîtrise des discours. Elle se fonde donc sur la prise en compte des situations de communication. » *ibid.* ;
« Pour ce qui concerne les catégories grammaticales, il convient de procéder à une révision et à un approfondissement systématique des notions fondamentales en grammaire de phrase, en grammaire de texte et en grammaire du discours.
Les développements contemporains de la pragmatique du discours (énoncé, énonciateur, situation d'énonciation, implicite et sous-entendu, etc.) nécessitent un apport de formation tout particulier. », *Rapport officiel du jury Capes interne et Caer de Lettres Modernes*, 2003.

[2] En témoigne également le colloque qui nous réunit : « Corpus en Lettres et Sciences sociales - Des documents numériques à l'interprétation ».

ampleur et un niveau de détail adaptés au degré de finesse et à la richesse attendue en résultat de l'analyse (*complétude*).

• *Des conditions d'exploitabilité* : Les textes qui forment le corpus doivent être commensurables (*homogénéité*). Le corpus doit apporter suffisamment d'éléments pour pouvoir repérer des comportements significatifs (au sens statistique du terme) (*volume*). »

Nous suivons F. Rastier (2005 : 32), qui propose la définition suivante du *corpus* (c'est nous qui soulignons) :

> *« Un corpus est un regroupement* ***structuré*** *de* ***textes intégraux****,* ***documentés****, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive* ***en tenant compte des discours et des genres****, et (ii) de manière pratique en vue d'une gamme d'applications*. »

Nos travaux s'inscrivent dans la perspective d'une linguistique textuelle capable de penser la diversité des textes, et privilégiant l'étude du sens textuel. Nous adoptons une conception praxéologique (praxéologie entendue comme *théorie de l'action dans et par le langage*) selon laquelle tout texte procède d'un genre, et tout genre est relatif à un discours (politique, littéraire, scientifique, etc.), attaché à une pratique sociale[1].

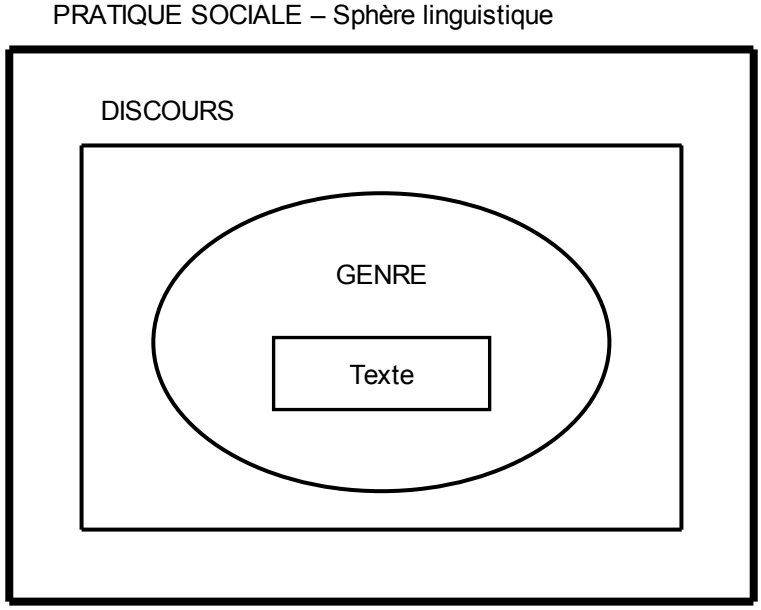


Fig. 3 : Conception praxéologique

Les textes sont ainsi étudiés dans leur production et leur interprétation, et sont reliés à leur entour social et historique.
Il s'agit notamment de déterminer l'incidence du genre textuel, du discours d'appartenance ainsi que de l'intertexte (corpus de référence) sur les unités linguistiques locales. Déterminations du global sur le local que l'on peut représenter ainsi :

[1] Cf. notamment Rastier F., 2001.

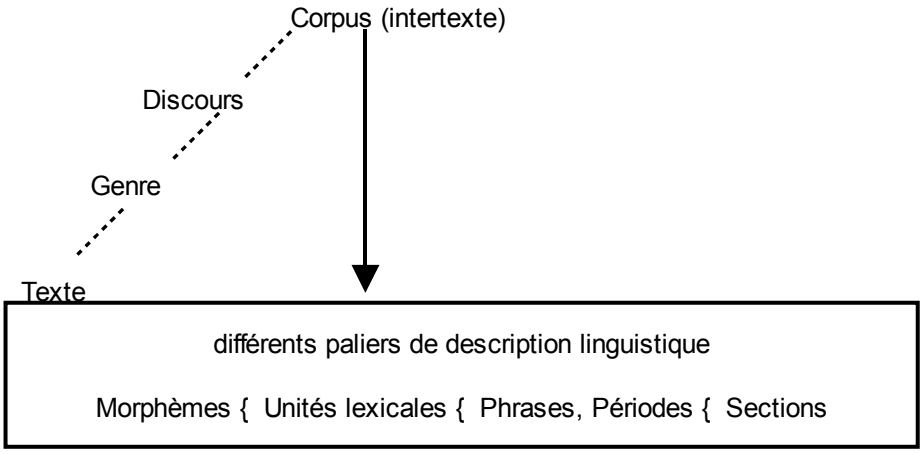


Fig. 4 : Déterminations du global sur le local (principe herméneutique)

2.2. Typologies textuelles

Comme les Traitements automatiques du langage ont affaire à des textes, non à des phrases, leur typologie est une condition de leur analyse. La diversité des genres et des discours amène à s'interroger sur l'importance de ces normes. Il apparaît ainsi nécessaire, dans la perspective d'établir des typologies textuelles[1], de tenir compte des genres et des discours.

L'étude des corpus montre par exemple que le lexique, la morphosyntaxe varient avec les genres et les discours. Denise Malrieu et François Rastier (2001) ont montré leur incidence sur l'ensemble des catégories morphosyntaxiques, ainsi que sur des variables comme la longueur des mots et des phrases. Aussi, disposer de critères et d'outils pour la classification des genres et des discours représente un enjeu pour la linguistique de corpus et la constitution de corpus pleinement utilisables (structuration des bases textuelles) pour des tâches de description linguistique.

Nous proposons de définir le genre par la cohésion d'un faisceau de critères linguistiques. Étiqueter le corpus avec des catégories morpho-syntaxiques ne suffit pas à faire émerger des genres. Les genres sont définis par des interactions normées entre diverses composantes et par des critères de corrélation entre ces composantes et leurs constituants :

lexico-syntaxiques, *sémantiques*, *sémantico-rhétoriques*

Les genres ne sont pas isolés mais contrastent au sein de champs génériques[2]. On ne caractérise donc pas les genres en soi mais par rapport aux autres genres, à commencer par les plus proches, à savoir ceux relevant d'un même champ générique[3].

Le schéma suivant illustre les relations intergenres au sein d'un même champ générique. Le sous-corpus 1 rassemble des textes appartenant au 'genre X'. Le sous-corpus 2 rassemble des textes appartenant au 'genre Y'. 'Genre X' et 'Genre Y' contrastent au sein d'un même //champ générique//, le champ générique A.

---

[1] Il ne s'agit pas de classer des types (fonctionnels) de textes ou d'isoler des séquences textuelles (narrative, descriptive, explicative, argumentative, etc.). Un texte ne se réduit pas à une étendue de plusieurs phrases ou propositions, et les unités textuelles ne sont pas nécessairement discrètes ; elles sont davantage à envisager comme des lieux et moments de parcours énonciatifs et interprétatifs. Cf. Rastier F., 2003.

[2] Cf. Rastier F., 2001, Glossaire p. 297.

« *Champ générique* : groupe de genres qui contrastent voire rivalisent dans un champ pratique : par exemple, au sein du discours littéraire, le champ générique du théâtre se divisait en comédie et tragédie ; au sein du discours juridique, les genres oraux constituent un champ générique propre (réquisitoire, plaidoirie, sentence). ».

[3] Le regroupement des genres en champs génériques fait appel à des critères situationnels (comme le contexte de production/de réception, les objectifs et visées, le support médiatique, le mode de diffusion, etc.). Chaque champ générique correspond à un champ pratique ; une connaissance de la pratique sociale paraît nécessaire pour effectuer les regroupements.

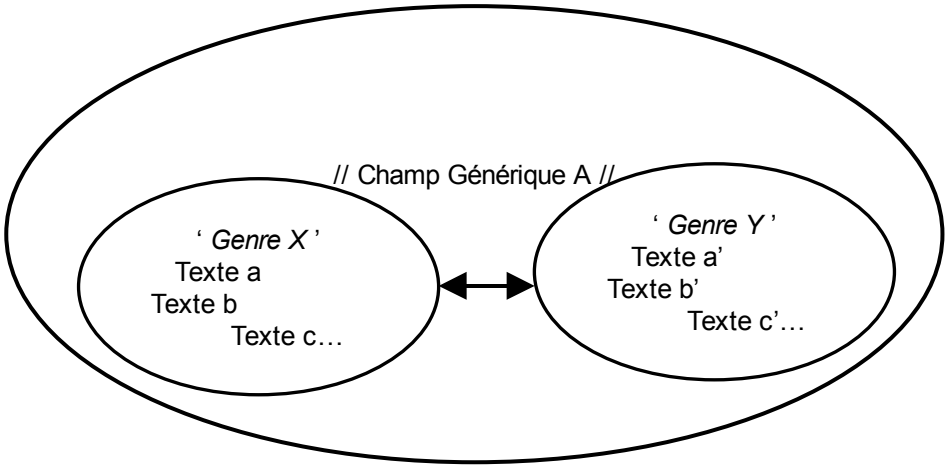


Fig. 5 : Relations contrastives entre genres au sein d’un même champ générique

Les champs génériques sont ainsi envisagés comme des classes d’interdéfinition des genres. Il s’agit alors de proposer une caractérisation différentielle des genres et d’établir les traits spécifiques de chacun d’eux.
Nous faisons l’hypothèse que les procédés sémantico-rhétoriques[1] jouent un rôle important dans la caractérisation des genres. Pour permettre leur traitement, nous cherchons actuellement à identifier des marqueurs de procédés sémantico-rhétoriques et à automatiser leur balisage.

Conclusion

L’étude du statut des « groupements de textes » et de la place qu’ils occupent dans l’enseignement du français (dans le secondaire) nous a conduit à problématiser la notion de « corpus » en linguistique et à proposer des critères pour la caractérisation des genres et la typologie des textes.
En lien avec cette problématique, une étude sur l’exploitation des corpus numériques en classe mériterait d’être menée. Il s’agirait d’analyser les apports du travail sur corpus et de s’interroger sur les applications qui peuvent être proposées : recherche par mots-clés ou mots-vedettes ? ; étude de corrélations ? ; avec quels outils et suivant quels objectifs ?
On pourra notamment s’appuyer sur les nombreuses analyses de corpus réalisées par Thierry Mézaille, dans la perspective d’une *sémantique textuelle, littéraire et didactique*[2].

BIBLIOGRAPHIE

ADAM, J.-M. 1992. *Les textes : types et prototypes*, Paris, Nathan.
ADAM, J.-M. 1999. *Linguistique textuelle. Des genres de discours aux textes*, Paris, Nathan.
ARMAND, A., DESCOTES, M. et al. 1992. *La séquence didactique en français : classes de lycée, CAPES, agrégation*, Paris, Bertrand Lacoste.
BIARD, J. & DENIS, F. 1993. *Didactique du texte littéraire. Progressions et séquences*, Paris, Nathan.
BIBER, D. 1988. *Variations across Speech and Writing*, Cambridge, Cambridge University Press.
BIBER, D. 1995. *Dimensions of register variation : a cross-linguistic comparison*. Cambridge University Press, Cambridge.
BOMMIER-PINCEMIN, B. 1999. *Diffusion ciblée automatique d’informations : conception et mise en œuvre d’une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de l’Université Paris IV, 805 p.
BRONCKART, J.-P. 1985. *Le fonctionnement des discours*, Neuchâtel ; Paris, Delachaux et Niestlé.
COMBETTES, B. 2002. Analyse linguistique des textes et stylistique, *Langue française*, 135, pp. 95-113.

[1] Les procédés sémantico-rhétoriques renvoient notamment aux composantes de l’*invention-élocution* (composantes éthique, argumentative, pathétique). Pour une présentation, cf. Duteil-Mougel C., 2005.
[2] Cf. le site personnel de l’auteur : http://www.chez.com/mezaille/.

CONDAMINES, A. (dir.) 2005. *Sémantique et corpus*, Paris, Hermès.
DUTEIL-MOUGEL C. 2005. Les mécanismes persuasifs des textes politiques. Propositions théoriques pour l'analyse de corpus, *Corpus*, 4, pp. 129-155.
HABERT, B. 2006. Portrait de linguiste(s) à l'instrument, *Texto ! Textes et cultures*, Vol. X, n°4.
HABERT, B., NAZARENKO, A. & SALEM, A. 1997. *Les linguistiques de corpus*, Paris, Armand Colin - Masson.
HABERT, B. et coll. 2000. Profilage de textes : cadre de travail et expérience, in *Actes des 5èmes JADT*.
IDE, N. et VÉRONIS, J. 1995. *Text Encoding Initiative - Background and Context*, Dordrecht (The Netherlands), Kluwer Academic Publishers.
ILLOUZ, G., HABERT, B. et al. 1999. Maîtriser les déluges de données hétérogènes, in A. Condamines et al. (éds.), *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, Actes de l'atelier thématique TALN, Cargèse, pp. 37-46.
JORDY, J. 1995. Groupement de textes poétiques – Une séquence autour du blason, in M. Descotes et al., *Lire méthodiquement des textes*, Paris, Bertrand Lacoste.
JORDY, J. 1991. *Le groupement de textes*, Toulouse, CRDP.
Le Français aujourd'hui, « Le groupement de textes », n°97, Mars 1992.
MALRIEU, D. & RASTIER, F. 2001, Genres et variations morphosyntaxiques, *Traitement Automatique des langues*, vol. 42, 2, pp. 548-577.
MAYAFFRE, D. 2002. Les corpus réflexifs : entre architextualité et intertextualité, *Corpus*, I, 1, pp. 51-70.
MEZAILLE, T. 2003. *Thématiques littéraires - enseignement des textes numériques pour un accès sémantique et didactique aux banques textuelles*, HDR.
RASTIER, F. 2001. *Arts et sciences du texte*, Paris, PUF.
RASTIER, F. 2003. Parcours de production et d'interprétation, in A. Ouattara (éd.), *Parcours énonciatifs et parcours interprétatifs*, Ophrys, pp. 221-242.
RASTIER, F. 2005. Enjeux épistémologiques de la linguistique de corpus, in G. Williams (éd.), *La linguistique de corpus*, Rennes, PU Rennes, pp. 31-45.
RASTIER, F. à paraître. La traduction : interprétation et genèse du sens.
SUEUR, J.-P. 1982. Pour une grammaire du discours, *Mots*, 5, pp. 145-185.
VECK, B. 1995. *Groupements de texte et projet de lecture*, (2 vol.), Paris, Bertrand Lacoste.

Documents officiels :
***Programmes et accompagnement* (enseigner au collège, Français), 2004 (rééd.), Paris, Centre National de Documentation Pédagogique.
*Organisation des enseignements dans les classes de 6ème des collèges : Arrêté du 29 mai 1996 – (*BO* n° 25 du 20 juin 1996), modifié par l'arrêté du 14 janvier 2002 (*BO* n° 8 du 21 février 2002) ;Arrêté du 22 novembre 1995 relatif aux programmes de la classe de 6ème des collèges.
*Programme du cycle central des collèges : Arrêté du 10 janvier 1997. *JO* du 21 janvier 1997 – (*BO* Hors série n° 1 du 13 février 1997).
*Programme de la classe de 3ème, BOHS n° 10 du 15 octobre 1998.

***Français (classe de seconde),* 2003 (rééd.), Paris, Centre National de Documentation Pédagogique.
*Programme de la classe de seconde générale et technologique.
BO n° 41 du 7 novembre 2002.

***Français (classe de première), Littérature (classe terminale série littéraire)*, 2004 (rééd.), Paris, Centre National de Documentation Pédagogique.
*Programme de français des classes de première des séries générales et technologiques.
BO n° 28 du 12 juillet 2001.

**Epreuves anticipées de français
Note de service n° 2001-117 du 20 juin 2001. BO n° 26 du 28 juin 2001.
Modifiée par la note de service n° 2003-002 du 8 janvier 2003. BO n° 3 du 16 janvier 2003.

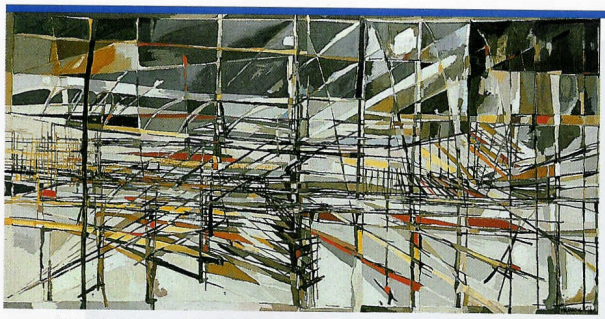
**Epreuve d'admissibilité au Capes interne de Lettres Modernes
Arrêté du 2 mars 2000 modifiant l'arrêté du 30 avril 1991 modifié fixant les sections et les modalités d'organisation des concours du certificat d'aptitude au professorat de l'enseignement du second degré.

**Rapport officiel du jury Capes interne et Caer de Lettres Modernes, 2005, présenté par Madame Catherine BECCHETTI-BIZOT, Inspectrice générale de l'éducation nationale, Présidente du jury, Paris, Centre National de Documentation Pédagogique.

**Rapport officiel du jury Capes interne et Caer de Lettres Modernes, 2004, présenté par Madame Catherine BECCHETTI-BIZOT, Inspectrice générale de l'éducation nationale, Présidente du jury, Paris, Centre National de Documentation Pédagogique.

**Rapport officiel du jury Capes interne et Caer de Lettres Modernes, 2003, présenté par Madame Catherine BIZOT, Inspectrice générale de l'éducation nationale, Présidente du jury, Paris, Centre National de Documentation Pédagogique.

ANNEXE
Extrait d'un manuel de Français pour la classe de troisième, Hatier, 1999, p.20



Vieira Da Silva (1908-1992), *Composition*, 1951 (huile sur toile; Bâle, Kuntmuseum, Offentliche Kunstsammlung).

RACONTER À LA PREMIÈRE PERSONNE

Leçon p. 54

J'écris cette histoire

TEXTE 1

Robert Merle, écrivain français, né en 1908.

Madrapour, roman, 1976.

Il s'agit ici du début du roman.

13 novembre.

J'écris cette histoire en même temps que je la vis. De jour en jour. Ou plutôt – ne soyons donc pas si ambitieux – d'heure en heure. Nous serions bien avisés, d'ailleurs, de faire tenir un monde dans chaque minute qui passe. Nous n'en avons pas tant à notre disposition. La vie la plus longue peut se chiffrer en secondes. Faites le calcul : c'est un chiffre qui n'a rien d'astronomique – ni de particulièrement rassurant.

Tandis que j'écris ceci, je suis bien incapable de prévoir la fin de mon aventure. Je ne pénètre pas non plus sa signification. Ce n'est pourtant pas faute de hasarder des hypothèses.

Mon histoire aura sûrement un terme, ne serait-ce que le plus évident. Mais il n'est pas certain qu'elle ait un sens ou – ce qui revient au même – il n'est pas sûr que je sois capable de lui en trouver un : « Un moucheron qui naît à l'aube et meurt au coucher du soleil ne peut pas comprendre le mot nuit. »

Quand le taxi me dépose à l'aéroport de Roissy-en-France, une surprise m'attend. Tout est vide. Ni voyageurs, ni employés, ni hôtesses. Je suis seul, rigoureusement seul, dans ce monument de métal et de glace, où règne un silence de crypte. Comparaison absurde : Roissy ressemblerait plutôt, avec ses immenses vitres, à une serre démesurée.

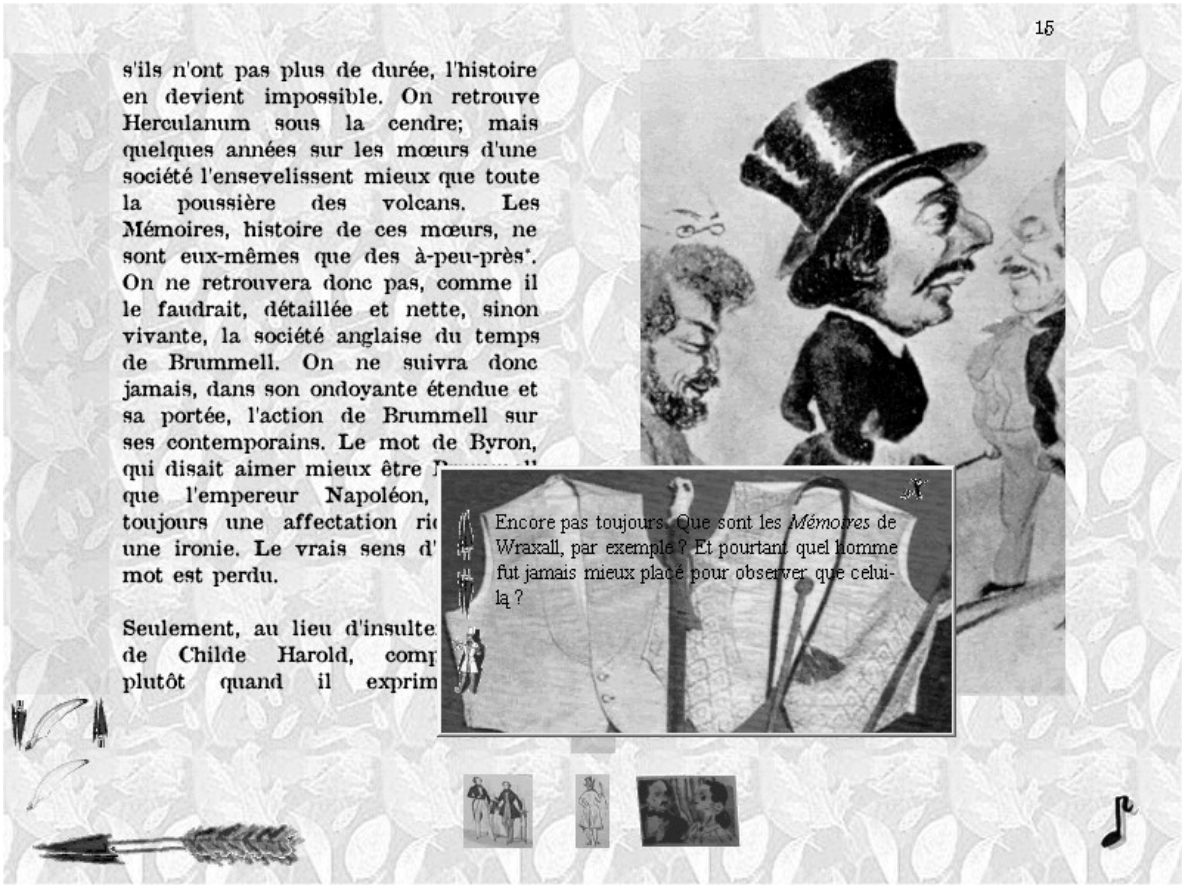
Je pose mes valises sur un chariot et, dans ce désert lumineux, je pousse le chariot devant moi. En même temps, je sens tout le ridicule de convoyer ainsi mes biens terrestres, alors qu'aucun préposé ne peut les prendre en charge.

Question posée aux élèves :
« Lire l'image
Observez l'ensemble des lignes, des figures, les tonalités dominantes dans le tableau de la page 20. Quel lieu ce tableau peut-il évoquer ? Citez une phrase du texte qui pourrait servir de légende à cette peinture. »

L'ŒUVRE COMPLÈTE NUMÉRIQUE DE BARBEY D'AUREVILLY

David COCKSEY
Université de Toulouse II / CUFR J-F Champollion

SOMMAIRE



1. Introduction

Lorsqu'en 2000 nous avons jeté les bases du travail dont il sera question ici, la publication au format numérique d'œuvres littéraires dites classiques ne constituait déjà plus une nouveauté. La société Acamédia avait récemment commercialisé *Alexandre Dumas, un aventurier de génie* (1997) puis *Chateaubriand, les itinéraires du romantisme* (1998) et *Balzac : explorer La Comédie humaine* (1999), réalisé avec l'appui du groupe international de recherches balzaciennes. À la même époque, Bibliopolis fournit des éditions numériques plus rudimentaires des romans de Zola et des frères Goncourt, entre autres, tandis que Gallimard plongea le lecteur dans l'esprit OuLiPo grâce aux *Machines à écrire*[1] de Bernard Magné.

C'est après avoir pris connaissance de certains de ces travaux que nous conçûmes l'idée d'une première approche assistée par ordinateur de l'œuvre aurevillienne. Notre projet réunit l'œuvre romanesque et critique de Barbey d'Aurevilly, soit 38 volumes[2] d'articles littéraires, historiques et

[1] *Cf.* http://www.ac-nantes.fr/peda/disc/lettres/ressourc/coeur/oulipo/oulipo.htm
[2] Les textes numérisés sont ceux de la première édition des *Œuvres et les Hommes*.

politiques ainsi que 15 romans et nouvelles[1] présentés à travers une interface graphique évocatrice de l'univers aurevillien. Conçu au départ comme un outil de travail pour notre thèse[2], il a évolué dans le cadre d'une réflexion sur la particularité de l'objet littéraire[3] multimédia : dans le domaine de textes anciens, quels sont les atouts de l'édition numérique, et quel rapport entretient-elle avec le livre classique ?

2. Objectifs du projet

La résurrection critique de Barbey romancier date des années 1960, mais celle du journaliste commence à peine. Ce décalage est regrettable : la dissociation du critique et du romancier qui en est issue risque de faire méconnaître l'un et l'autre. Pour mieux aborder l'œuvre aurevillienne comme un tout stylistique, il nous a paru souhaitable de présenter les deux volets de l'œuvre sur un même support. Seule une édition électronique pouvait permettre cette approche holistique, qui restitue l'unité de l'œuvre et donc de l'écrivain. Pour la première fois, romancier et critique se trouvent réunis, associés au dandysme qui les préside tous deux. Ce rapprochement des corpus met en évidence les passerelles génériques entre eux, et fait s'entendre une voix unie qui contribue à atténuer l'opposition longtemps perçue entre les deux volets de l'œuvre.

On a dit que Barbey était « un palais dans un labyrinthe[4] », et il en va de même pour son œuvre. Au niveau le plus élémentaire, le numérique doit faciliter l'accès à l'œuvre aurevillienne, en particulier à son volet critique. À l'heure actuelle, seule l'édition de la Pléiade propose l'intégralité de ses romans ; la quasi-totalité des 1300 articles de l'œuvre critique a été reprise en recueil entre 1861 et 1909. Si ces tomes désormais rares font actuellement l'objet d'une réédition[5], le volume ainsi que l'hétérogénéité de l'œuvre journalistique aurevillienne demeurent une barrière à leur exploitation efficace. Il n'est pas aisé de se repérer dans trente-six fois quatre cents pages d'acrobaties intellectuelles, d'autant plus que le classement thématique de ces recueils est nécessairement approximatif. En délimitant un corpus de travail au sein de l'œuvre, on s'expose donc à des oublis ainsi qu'à des choix involontairement arbitraires. Ainsi, l'accès à ce précieux document historique, littéraire et social demeure l'apanage des spécialistes de Barbey.

Or, Michel Butor remarque que le mode de lecture intégral, qui commence à la première page d'un livre pour finir à la dernière, n'est ni le seul ni même le plus répandu : les livres sont plutôt

> des réserves de savoir dans lesquels nous pouvons puiser, et qui sont arrangées de telle sorte que nous puissions trouver le plus facilement possible le renseignement dont nous avons besoin à un moment donné[6].

Dans cette optique, l'apport du livre électronique est évident : une fois numérisée, l'œuvre dévoile ses secrets à travers la recherche par mot-clé et selon des critères de classement plus intuitifs. Le spécialiste n'est plus exposé aux oublis et aux contre-sens, tandis que le curieux est en mesure de se renseigner à souhait et à la source sur l'avis de Barbey concernant telle ou telle question.

Sur le plan de la conception, notre projet dépasse sa visée première, encyclopédique, pour aboutir à une union réfléchie de l'écrit et de l'image. Sa dimension pluridisciplinaire se reflète dans les utilisations auxquelles il se prête. En faculté de Lettres, aurevilliens, linguistes, stylisticiens et spécialistes du rapport texte/image ou encore du multimédia pourront y avoir recours dans le cadre de leur recherche ou de leurs enseignements. En outre, l'œuvre critique aurevillienne s'ouvrira ainsi aux historiens de la littérature, de l'art, des idées, du journalisme, et de la politique.

Cette édition pourra également contribuer à une meilleure représentation de Barbey dans l'enseignement secondaire. Pour l'instant, les manuels présentent tant bien que mal quelques extraits de ses romans les mieux connus, mais il est difficile d'apprivoiser ainsi l'univers aurevillien. À l'heure des parcours diversifiés, on peut envisager que l'image, en

---

[1] Les éditions numérisées correspondent à celles à partir desquelles a été constituée celle de la Bibliothèque de la Pléiade (la première édition de chaque œuvre), à l'exception des *Memoranda*, numérisés à partir de l'édition Bernouard (1927). La seule version complète de ces cinq textes figurant dans la Pléiade, nous ne pouvons la reproduire.

[2] *Le Dandysme littéraire de Barbey d'Aurevilly*, University of Sheffield, 2004.

[3] *Cf*. Dominique Maingueneau, *Le Contexte de l'œuvre littéraire. Énonciation, écrivain, société*, Paris, Dunod, 1993, et Michel Butor, *Essais sur le roman*, Paris, Gallimard, coll. « Idées », 1972.

[4] Jules Barbey d'Aurevilly, *Du Dandysme et de George Brummell*, *Œuvres complètes*, Gallimard, « Bibliothèque de la Pléiade », II, 1966, p. 694. Le mot est d'Eugénie de Guérin.

[5] *Les Œuvres et les Hommes*, Paris, Les Belles Lettres, 2004-2010.

[6] Michel Butor, *op. cit.*, p. 138.

fournissant un accès alternatif à l'œuvre, facilite l'étude des textes par la suite[1]. En outre, il serait utile de mettre à la portée de l'élève à la découverte de son patrimoine littéraire les écrits de celui qui fut l'un des premiers défenseurs de Balzac et de Baudelaire ainsi que l'éreinteur semi-lucide de Hugo, de Flaubert et de Zola.

3. Un objet littéraire multimédia

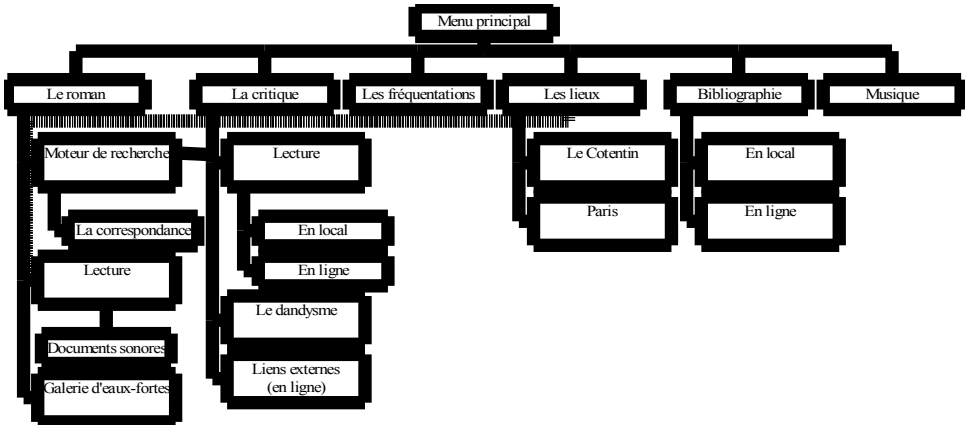
La réédition numérique de livres anciens invite à réfléchir sur la présentation du texte en tant que paramètre de la réception. Si le public se montre toujours réticent quant à la lecture à l'écran, c'est peut-être en partie parce que l'édition électronique rompt souvent avec les conventions de présentation propres au livre. Le support physique, lieu de rencontre entre auteur, texte et lecteur, est un détail évident mais non négligeable de la transmission du texte. Dans la mesure du possible, notre édition imite donc les supports d'origine, pour évoquer les conditions de réception de l'époque.

L'écrivain étant, comme on le sait, inséparable de son contexte historique, il peut s'avérer utile d'assister l'imagination du lecteur en faisant revivre ce contexte à travers des éléments iconiques. Maints endroits célèbres à l'époque, donc qui ont une valeur référentielle dans l'écriture, sont plus ou moins oubliés aujourd'hui. C'est le cas, par exemple, du café Tortoni ou de la Maison d'Or. Le lecteur moderne est donc privé de repères dont disposait son homologue de l'époque. De même, des décennies de vulgarisation ont fini par brouiller l'image du dandy. Nous avons donc reproduit des gravures d'époque, que l'on peut consulter ou bien les unes à la suite des autres, ou bien au fil de la lecture grâce à des liens hypertextes. Plus qu'une simple juxtaposition de l'image et du texte, il s'agit d'un dispositif où la présence de l'un participe de la réception de l'autre. Dans le même esprit, l'inclusion des eaux-fortes d'époque de Félicien Rops et surtout de Félix Buhot, compatriote normand de Barbey, donnent d'autres perspectives sur l'imaginaire aurevillien. En effet, ces gravures, qui représentent des passages-clés des récits concernés, où convergent logiques discursive et iconique[2], invitent à considérer la scène romanesque telle qu'elle se manifeste chez Barbey, perspective d'autant plus appropriée que *L'Ensorcelée* fut l'un des premiers romans français à donner lieu à une suite d'illustrations photographiques[3].

4. Réalisation initiale

Dès 2001, le projet a donné lieu à un cédérom agencé comme suit.

Arborescence



Menu principal. Cet écran permet d'accéder aux trois parties thématiques : le roman, la critique et l'homme, ce dernier vu à travers ses fréquentations ainsi que les lieux qu'il a habités et qui ont alimenté son imaginaire. À partir de cet écran, le lecteur a également la possibilité d'accéder à des fichiers d'aide en ligne. Les éléments du menu s'affichent en « Connétable », une police de caractères que nous avons développée à partir de reproductions de manuscrits aurevilliens. De type TTF, elle est compatible avec toute application Windows (Word, Internet Explorer, etc.).

---

[1] Académie de Montpellier, « Logiciels, cédéroms », http://pedagogie.ac-montpellier.fr/Disciplines/lettres/logiciels/acamedia/acamedia.htm.

[2] Cf. Stéphane Lojkine, *La Scène de roman*, Paris, A. Colin, 2002, p. 10.

[3] David Cocksey, « Henri Magron, photographe d'inspiration littéraire », *Histoires littéraires*, n°19, 2004.

Le roman. Cette partie comprend un écran de lecture, un moteur de recherche plein-texte et une galerie d'eaux-fortes.

Le dispositif de lecture imite le livre : le lecteur tourne des pages, auxquelles on peut associer des annotations en marge. On peut également placer des marque-pages. La page en cours peut s'imprimer avec les annotations qui l'accompagnent. Le support physique d'un texte étant le relais entre l'auteur et le lecteur, un objet d'art qui participe de la réception du texte, nous utilisons une police de caractères ancienne, évocatrice à la fois du livre de l'époque et des journaux, où on a pu lire pour la première fois la plupart des textes de Barbey. L'impact du support moderne sur le texte ancien se trouve ainsi minimisé.

Le moteur de recherche gère les opérateurs booléens ainsi que les *wildcards*[1], affichant le nombre d'occurrences par œuvre. La dernière version du moteur comporte un historique et permet de rechercher dans plus d'un corpus à la fois ainsi que de passer plus commodément d'un corpus à un autre. En surlignant à l'aide de la souris un passage du texte affiché, le lecteur le récupère dans un fichier Word créé dans C:\Mes documents.

La galerie d'eaux-fortes rassemble des œuvres de Félix Buhot, de Félicien Rops et de Lobel-Riche.

La critique.
Le **moteur de recherche** montre ici tout son potentiel en traitant l'ensemble des 1300 articles de Barbey.

L'écran de lecture permettra dans la version définitive de consulter les articles par date, par journal de parution ou par recueil de réédition[2]. Ce premier classement fidèle de l'ensemble des articles permettra de mesurer l'évolution dans le temps de l'écriture critique de Barbey ainsi que l'éventuel impact sur les articles des différents journaux auxquels ils étaient destinés. La présentation du texte en trois colonnes renoue avec la mise en page journalistique, et rapproche le lecteur de l'objet littéraire original.

L'écran internet permet d'accéder à des livres ou à des œuvres d'art évoqués par Barbey dans ses articles. Moins autonome que le roman ou la poésie, l'écriture critique ne s'apprécie pleinement qu'à travers une certaine connaissance des œuvres auxquelles elle fait référence ; cette mise en contexte est facilitée par le multimédia. Le cédérom renvoie donc à une page web susceptible d'évoluer pour constituer un portail dix-neuviémiste. Ainsi, le lecteur peut retrouver le paysage littéraire de l'époque, très différent de celui de la postérité.

Du Dandysme et de George Brummell occupe une place unique au sein de l'œuvre aurevillienne, à la frontière de la critique et du romanesque. Nous avons donc choisi de le présenter à part, d'autant plus que le dandysme appelle une mise en contexte. Le texte de Barbey est plutôt abstrait, et présuppose chez le lecteur des notions élémentaires sur le sujet. Or l'évolution des usages et des mentalités fait qu'on ne parvient plus à se représenter ce type par définition éphémère. Le support iconographique laisse entrevoir la réalité historique des dandys, éclipsée aujourd'hui par l'acception vulgarisée du mot. Le dispositif de lecture comporte une spécificité : il s'agit d'une fenêtre, destinée à l'affichage des notes en bas de page, qui représentent un tiers de l'ouvrage. Ainsi se traduit à l'écran la rupture avec la lecture linéaire, technique typique de l'écriture dandy.

Paris. À terme, cet écran recréera le milieu du journaliste et du dandy, tout comme celui consacré à **la Normandie** permet au lecteur de retrouver le berceau du romancier en visionnant des paysages qui font l'objet d'une mise en scène aussi fantastique que réaliste. Des liens hypertextes mettent ou mettront l'iconographie en relation directe avec les textes concernés.

---

[1] Recherche troncature : journa*, par exemple, pour trouver toutes les occurrences de journal, journaux, journalisme et journaliste(s).

[2] En attendant la constitution d'une véritable base de données capable de gérer ces modes de tri, seul le moteur de recherche permet d'accéder à l'ensemble des textes.

Les Fréquentations. Ouverture sur le Barbey des écrits intimes, cette partie complète le panorama des trois registres de l'écriture aurevillienne. La diversité de Barbey et de ses fréquentations apparaît dans sa correspondance, dont on lira un échantillon classé par ordre chronologique ou par correspondant. L'iconographie permettra de donner corps aux voix épistolaires. Le réseau littéraire fin-de-siècle qui se constitua autour du « Connétable de Lettres » sera présenté de la même manière.

5. État d'avancement

La maquette a été réalisée sous Macromedia Director, logiciel de programmation orienté objet destiné à la création de cédéroms. Or, si Director permet une excellente gestion de l'interface graphique, la manipulation de données textuelles n'est pas sa visée première. Du reste, une parution sur cédérom n'aurait pas été sans inconvénients : se seraient posées des questions de compatibilité PC/Macintosh, de compatibilité ascendante[1], et de diffusion restreinte (le sort de Bibliopolis et d'Acamédia n'a pas été des meilleurs) : autant de considérations problématiques pour un projet destiné à améliorer la visibilité de l'œuvre de Barbey. Depuis 2002, nous envisageons donc de transformer le cédérom en site web, choix d'autant plus facile que le développement récent du haut débit permet de proposer des sites de plus en plus élaborés. Notre première maquette mise en ligne propose une centaine d'articles selon les mêmes critères de tri prévus pour le cédérom. Réalisé en HTML avec un moteur de recherche PHP, il est fonctionnel au point de connaître un certain succès[2] ; mais il est devenu apparent que cette architecture n'est pas suffisamment performante pour intégrer l'ensemble des textes. Nous avons donc décidé, avec le soutien technique de la Mission Multimédia de l'Université de Toulouse II, de refonder le site en nous appuyant sur une architecture XML. Ce travail, qui implique un balisage minutieux des textes, a débuté au mois d'avril.

Nous aurons l'occasion lors du colloque de revenir plus amplement sur certains des thèmes abordés ci-dessus, tout en en développant d'autres, dans le cadre d'une présentation des trois maquettes réalisées.

BIBLIOGRAPHIE

ACADÉMIE DE MONTPELLIER. « Logiciels, cédéroms »,
http://pedagogie.ac-montpellier.fr/Disciplines/lettres/logiciels/acamedia/acamedia.htm

BALZAC : explorer La Comédie humaine, cédérom, Paris, Acamédia, 1999.

BARBEY D'AUREVILLY, J. 1964, 1966. *Œuvres complètes*, Gallimard, Bibliothèque de la Pléiade, 2 vols.

BARBEY D'AUREVILLY, J. *L'Œuvre complète numérique*, site internet en cours de réalisation sous la direction de D. Cocksey : http://www.univ-tlse2.fr/lla/barbey/oc/

BUTOR, M. 1972. *Essais sur le roman*, Paris, Gallimard, coll. « Idées ».

CHATEAUBRIAND 1998. *Les itinéraires du romantisme*, cédérom, Paris, Acamédia.

COCKSEY, D. 2004. *Le Dandysme littéraire de Barbey d'Aurevilly*, Thèse de doctorat, University of Sheffield,

COCKSEY, D. 2004. Henri Magron, photographe d'inspiration littéraire, *Histoires littéraires*,19.

Alexandre DUMAS, un aventurier de génie, cédérom, Paris, Acamédia, 1997.

GIDE, A. 2001. *Les Caves du Vatican*. Édition génétique, conçue et présentée par Alain Goulet et réalisée par Pascal Mercier, cédérom, Gallimard et université de Sheffield. *Cf.* http://www.gidiana.net/CD-ROM_Genetique.html.

LOJKINE, S. 2002. *La Scène de roman*, Paris, A. Colin.

MAGNÉ, B. & DENIZE, A. *Machines à écrire*, cédérom, Gallimard. *Cf.* http://www.ac-nantes.fr/peda/disc/lettres/ressourc/coeur/oulipo/oulipo.htm

MAINGUENEAU, D. 1993. *Le Contexte de l'œuvre littéraire. Énonciation, écrivain, société*, Paris, Dunod.

[1] Compatibilité avec les systèmes d'exploitation ultérieurs.

[2] Répertorié par la BNF, l'Université de Paris X, l'Université de Lyon III et divers portails littéraires (la Société des Études romantiques et dix-neuviémistes, la Bibliothèque Virtuelle de la Philologie Romane, weblettres.net, alalettre.com…), il accueille, d'après le Centre Interuniversitaire de Calcul de Toulouse qui l'héberge, un millier d'internautes par mois.

L'EMPLOI DES MÉTHODES DE LA LINGUISTIQUE DE CORPUS POUR L'ATTRIBUTION DE TEXTES : LES CARACTÉRISTIQUES CONCEPTUELLES, SÉMANTIQUES ET ÉPISTÉMOLOGIQUES DU LEXÈME *FAITH* DANS LES TEXTES DE SHAKESPEARE

Natalia N. BELOZEROVA
Université d'Etat de Tumen (Russie)

Dans cet exposé nous tenterons de montrer le rôle des bibliothèques électroniques dans la constitution d'un corpus linguistique, son traitement primaire et son exploitation.
Le sous-titre de la communication indique les approches, les catégories linguistiques, les conceptions et les méthodes d'étude, élaborées et argumentées sur la base de textes différents faisant partie des bibliothèques électroniques, y compris les méthodes très usitées comme celles des analyses comparative, définitionnelle, de constituants, structurale, conceptuelle et discursive. C'est sur leur base que s'élabore la méthode de l'attribution d'un texte à un auteur, méthode qui permet d'aller jusqu'à la formulation d'hypothèses d'inversions conceptuelles. Parmi les catégories illustrées sur la base des corpus de bibliothèques électroniques, nous proposons et développons les catégories de l'*intertextualité*, de l'*hypertextualité* et de la *fractalité* ; nous développerons également la catégorie modifiante de l'*interexistencialité*, qui s'est imposée dans le développement logique de notre recherche.
Quand nous parlons de l'utilisation de bibliothèques électroniques, nous empiétons involontairement sur le domaine de la linguistique de corpus. Ce courant prit forme et développa son argumentation théorique et son statut simultanément avec le développement d'Internet. Les premiers travaux, écrits dans le cadre de la linguistique de corpus, apparurent dans les années quatre-vingts, quand les grandes bibliothèques de l'Amérique et de l'Europe furent reliées pour la première fois par des réseaux d'ordinateurs[1]. On utilise la linguistique de corpus à des fins lexicographiques[2], pour l'analyse de textes et du discours[3], pour l'enseignement des langues et d'autres matières[4], pour la traduction[5], pour la rédaction des programmes de traduction automatique[6], et également pour des études comparatives[7].
Il nous faut également souligner le rôle particulier de la linguistique de corpus dans l'organisation des systèmes informatiques de recherche dans le cadre d'Internet, d'ordinateurs personnels et de certains porteurs de CD-Rom. Notons que les bibliothèques électroniques, dès les débuts de leur

---

[1] Voir; par ex., *Biber, D. 1988. Variation across speech and writing.* Cambridge: Cambridge University Press. Johns, T. 1988. *Whence and Whither Classroom Concordancing?* In T. Bongaerts et al (eds.) Computer Applications in Language Learning. Dordrecht*:* Foris. Klavans, Judith L. and Evelyne Tzoukermann. 1989. *Movement Verbs in English-French Translation: A Corpus-based Approach.* Proceedings of the Sixth Israeli Conference of Artificial Intelligence and Computer Vision. Tel Aviv, Israel. Sampson, G. 1987. *Evidence Against the``Grammatical"/`Ungrammatical" Distinction.* In W. Meijs (ed) Corpus Linguistics and Beyond. Amsterdam: Rodopi.

[2] La confection et la description de dictionnaires: Haslerud, V. & A-B. Stenstrom**.** 1995. *The Bergen Corpus of London Teenage Corpus (COLT)*. In G. Leech, G. Meyres & J. Thomas (eds) *Spoken English on computer*. London: Longman, 235-242).

[3] Andersen, G. & A-B. Stenstrom. *Forthcoming. A corpus-based investigation of the discourse items cos and innit. Synchronic Corpus Linguistics*. Toronto. Resnik, Philip. 1995. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. IJCAI-95.

[4] Jordan, G. 1992. *Concordances: Research Findings and Learner Processes.* Unpub M.A. Dissertation. Stevens, V. 1991. *Concordance-based Vocabulary Exercises: A Viable Alternative to Gap-Filling.* English Language Research Journal (4) 47--61. University of Birmingham.

[5] Par ex., Davis, Mark, Ted Dunning and Bill Ogden. 1995. *Text Alignment in the Real World: Improving Alignments of Noisy Translations Using Common Lexical Features, String Matching Strategies and N-Gram Comparisons.* European Association for Computation Linguistics*.*

[6] Klavans, Judith L. and Evelyne Tzoukermann. 1990. *Linking Bilingual Corpora and Machine Readable Dictionaries with the BICORD System.* Proceedings of the Sixth Conference of the University of Waterloo Centre for the New Oxford English Dictionary and Text Research: Electronic Text Research, University of Waterloo, Canada.

[7] Dunning, Ted, Jim Cowie and Takahiro Waka. 1991*. Analysis of Parallel Japanese and English Corpora.* CLR Tech Report. (MCCS-91-233).

création, se constituèrent à partir de la linguistique de corpus (d'Internet ou d'on-line et sur des porteurs de CD-Rom). Il convient d'ajouter des dictionnaires électroniques et des méthodes, proches des bibliothèques électroniques par leurs conceptions respectives, leur but et leur organisation.

La notion de **corpus** (*corpus, corpora*) est la notion centrale de ce courant. David Crystal en donne la définition suivante : « Le **corpus**, pl. **corpora** – la collection (l'ensemble, la réunion) de données linguistiques, constituée soit comme des textes écrits, soit comme la transcription du langage parlé. Le but principal de tout corpus de données est l'argumentation d'une hypothèse quelconque de la nature ou du fonctionnement de la langue, par exemple, définir, comment l'emploi d'un certain son, d'un mot ou d'une construction syntaxique varie. Les principes et l'emploi concret de **corpora** (corpus de données linguistiques) pendant l'étude linguistique sont l'objet de la **linguistique de corpus**. **Le corpus électronique** est une grande réunion de textes, notés pour la lecture automatique » (cf. Crystal, David. 1992. *An Encyclopedic Dictionary of Language and Languages*. Oxford, 85).
Nous voyons que ce dictionnaire considère le corpus comme l'ensemble ou la collection de données linguistiques nécessaires pour l'étude de la langue, et qu'il différencie le corpus linguistique et le corpus électronique. La linguistique de corpus, selon la définition présente, ne se focalise pas sur la description du système de la langue, mais sur les principes de l'emploi concret d'un corpus de données.
Envisageons la deuxième définition :
« **Corpus** (XIII-e siècle du mot latin *corpus*, le corps. Le pluriel est d'habitude *corpora). (1)Toute réunion de textes, particulièrement complet et indépendant des autres réunions de textes, par exemple, le corpus de la poésie anglo-saxone: (2)Le pluriel est aussi corpus. Dans la linguistique et la lexicographie – toute réunion de textes, d'énonciations ou d'autres exemples linguistiques qu'on conserve comme une base électronique de données. A présent, les corpus électroniques (corpora) peuvent conserver des millions de mots employés et vieillis, dont les propriétés peuvent être analysées à l'aide de classifications et de catégorisations (tagging - l'addition aux mots et à d'autres formations d'explications définissantes et classifiantes), et aussi à l'aide de programmes de concordance* (concordancing programs). La linguistique de corpus étudie les données de tels corpus » (cf. McArthur, Tom "Corpus" , in: McArthur, Tom (ed.) 1992. *The Oxford Companion to the English Language*. Oxford, 265-266).

Nous voyons que le dictionnaire linguistique d'Oxford, lui, ne distingue pas le corpus linguistique et le corpus électronique. En outre, ce dictionnaire accentue le caractère systématique de tout corpus de données et mentionne les méthodes principales de la linguistique de corpus. Si le sens du terme *tagging* se conçoit aisément d'après la définition citée, le terme *concordance* (concordancing programs) en revanche, exige des explications. Nous essaierons d'expliquer ce terme à partir de notre interprétation des termes CORPUS et CORPORA.
A notre avis, tout corpus de textes ou d'autres données (ou CORPORA, c'est-à-dire une grande accumulation de données) peut exister à plusieurs niveaux : (1) au niveau des documents en papier, y compris les bibliothèques traditionnelles, (2) au niveau virtuel (tel le niveau électronique dans l'ordinateur, l'Internet ou sur les CD, (3) au niveau idéal, comme l'information emmagasinée dans les cellules du cerveau d'individus, (4) au niveau de la Sémiosphère, y compris celle du subconscient collectif, (5) enfin au niveau de la Noosphère, réunissant tous les niveaux que nous venons d'énumérer. Dans ce cadre, la linguistique de corpus est contiguë à la fois à la sémiolinguistique, incluant la théorie de l'intertexte et de l'hypertexte, et à la linguistique d'ordinateur qui utilise la théorie de l'hypertexte et la théorie bibliothécaire, dans le cadre de laquelle les principes de catalogues et de références bibliographiques ont été élaborés. Ce sont ces principes ainsi que la théorie de l'intertexte, qui sont à la base des programmes de concordance (ou de programmes de recherche à l'intérieur de différents *corpora*, principalement des bibliothèques électroniques).
Les premiers auteurs des programmes de concordance partaient des thèses suivantes : (1) De tels programmes doivent devenir l'outil principal du chercheur dans le domaine de la linguistique de corpus. (2)Puisque la majeure partie des *corpora* est infiniment grande, il est inutile d'étudier tout corpus sans l'aide de l'ordinateur. (3) Le programme de concordance transforme les textes électroniques en une base de données qui peut être étudiée. (4) La recherche peut se réaliser

dans le cadre de mots isolés, de groupes de mots et aussi dans le cadre de morphèmes isolés (par ex., d'affixes concrets). (5) Si le programme est plus complexe, il peut contenir des listes nécessaires de groupes de mots et des listes des constituants les plus répandus[1].

Dans notre cas, pour illustrer les possibilités d'emploi des méthodes de la linguistique de corpus pour l'attribution de textes, il faut choisir un corpus idéal de textes.

A notre avis, l'œuvre de William Shakespeare en édition originale constitue un tel corpus idéal, des bibliothèques entières étant constituées d'ouvrages écrits sur le problème de l'attribution de cette oeuvre. Choisissons par exemple le soixante-sixième sonnet[2] et suivons sa corrélation de lexèmes (sa congruence) avec toute l'œuvre de Shakespeare. Si pendant l'analyse on découvre des corrélations sémantiques et épistémologiques, on peut considérer ce travail comme une contribution au procès de l'attribution du corpus qu'on étudie au poète, dramaturge, copropriétaire d'un théâtre, groom de la cour William Shakespeare, fils de John Shakespeare, un marchand, ayant obtenu le poste du chef de la ville de Stradford-sur-Avon et des armes nobles. Avec cela, on reconnaît le soixante-sixième sonnet comme une grandeur connue, puisqu'il existe une édition des sonnets de Shakespeare (1609), les mêmes sonnets étant entrés dans la première édition posthume de toute l'œuvre du poète (First Folio), recueillie par les amis du poète[3] qui avaient témoigné qu'il en était bien l'auteur. Les autres oeuvres, sauf deux poèmes, «Venus et Adonis» et «Lucrèce déshonorée»[4], qui avaient été publiés avec sa propre dédicace à son protecteur, le comte de Southhamton en 1593-1594, furent attribuées à un grand nombre de personnes célèbres et nobles (par ex., à Francis Bacon, Christopher Marlowe ou même à la reine Elisabeth I (Tudor)), d'où la question de la paternité de ces œuvres attribuées à Shakespeare jusqu'à preuve du contraire. Nous avons utilisé des moyens techniques permettant une recherche rapide et sûre des œuvres présentées par la bibliothèque électronique des oeuvres de Shakespeare et par le programme de recherche «concordance», pour démontrer l'unité et l'interexistencialité de tous les textes de Shakespeare.

Citons le texte du soixante-sixième sonnet :

1. Tir'd with all these, for restful ***death*** *I cry:*
2. As to behold ***desert*** *a* ***beggar*** *born,*
3. And needy ***nothing*** *trimm'd in* ***jollity****,*
4. And purest ***faith*** *unhappily forsworn,*
5. And gilded ***honor*** *shamefully misplac'd,*
6. And maiden ***virtue*** *rudely strumpeted,*
7. And right ***perfection*** *wrongfully disgrac'd,*
8. And ***strength*** *by limping sway disabled,*
9. And ***art*** *made tongue-tied by* ***authority****,*
10. And folly (doctor-like) controlling ***skill****,*
11. And simple ***truth*** *miscall'd* ***simplicity****,*
12. And captive ***good*** *attending* ***captain*** *ill:*
13. Tir'd with all these, from these would I be gone,
14. Save that to die, I leave my ***love*** *alone.*

(William Shakespeare : *Sonnets*, p. 67.)

---

[1] Les programmes les plus célèbres de concordance:

WordCruncher est un programme de concordance qui inclut une liste de fréquence *corpora* et de mots-clés dans leur entourage contextuel qui réalise la recherche de mots, de groupes de mots et de parties du mot. Ce programme figure sur le même CD que les *corpora* textuels (par exemple, le recueil de textes anglais médiévaux).

TACT (Text Analysis Computing Tools): un programme gratuit de concordance bien connu. Il a les mêmes fonctions que le **WordCruncher** et offre en complément la fonction de la recherche de collocations et de la démonstration comment un mot isolé fonctionne dans tout le corpus.

TACTWeb est un programme de concordance, fondé sur le programme TACT, mais destiné au travail dans le World Wide Web (voir TACTWebHomepage).

[2] Written between 1593 and 1609. First (unauthorized) edition: Shakespeare's Sonnets. Never before Imprinted, London, printed by George Eld, for Thomas Thorpe, 1609. Sonnets 138 and 144 were published earlier in "The Passionate Pilgrim" (1599).

[3] Ben Jonson, John Heminge and Henry Condell.

[4] Venus and Adonis,Composed 1592-1593. First printed by Richard Field, 1593 (apparently authorized text).
The Rape of Lucrece, Composed 1593-1594. First Edition: Lucrece (title-page). The Rape of Lucrece (heading of poem), printed by Richard Field, for John Harrison, 1594.

Pour la commodité de la description les lignes du sonnet sont numérotées. Dans tout le corpus, pour chacun des lexèmes choisis tous les contextes de l'emploi ont été extraits à l'aide d'un programme de recherche. Les significations des mots, l'étymologie, les dates d'entrée et ainsi que leur thésaurus ont été définis à l'aide de l'encyclopédie électronique en 32 volumes *Encyclopaedia Britannica 2003* (partiellement 2004) Deluxe Edition CD-ROM. Outre cela, nous avons eu recours à d'autres dictionnaires (voir la liste des dictionnaires), y compris à des dictionnaires on-line.
En tout 20 lexèmes ont été choisis (*death, desert, beggar, nothing, jollity, faith, honor, virtue, perfection, strength, art, tongue-tied, authority, doctor-like, skill, truth, simplicity, good, captain, love)*. Dans la communication présente nous allons nous limiter à l'analyse de l'emploi du lexème *faith*[1].
Prenons la quatrième ligne du 66-e sonnet «*And purest faith unhappily forsworn»* et essayons de définir les différents sens que Shakespeare avait construits à l'aide du mot *faith*.
On rencontre ce mot *faith* 448 fois dans les textes de Shakespeare. De ces emplois, 333 cas sont des groupes de mots prépositionnels *in faith (faith)* et *by (my) faith*. Dans les répliques des héros de Shakespeare ces expressions sont partiellement désémantisées, car elles transmettent les assurances de la véridicité de l'énonciation, correspondant en quelque sorte aux expressions françaises «ma parole!», «ma foi» et «crois-moi!». Citons quelques exemples :
1.LUCE
(Within.)
Faith, no, he comes too late,
[William Shakespeare: The Comedy of Errors,. 8463 (. Shakespeare-Riverside,. 91)]
2. TRA.
So could I*, faith,* boy, to have the next wish after,
That Lucentio indeed had Baptista's youngest daughter.
[William Shakespeare: The Taming of the Shrew, 8574 (Shakespeare-Riverside,. 116)]
3. KATH.
I' faith, sir, you shall never need to fear.
[William Shakespeare: The Taming of the Shrew, 8564 (. Shakespeare-Riverside, 114)]
4. PET.
You lie, *in faith,* for you are call'd plain Kate,
And bonny Kate, and sometimes Kate the curst;
[William Shakespeare: The Taming of the Shrew, 63.,. 8603 (Shakespeare-Riverside, 121)]
5. Now *by my faith and honor,*
If seriously I may convey my thoughts
In this my light deliverance, I have spoke

---

[1] Faith
Main Entry: 1faith
Pronunciation: ‚f†th
Function: noun
Inflected Form: pluralfaiths \‚f†ths, sometimes ‚f†Œz\
Etymology: Middle English feith, from Old French feid, foi, from Latin *fides*; akin to Latin fidere to trust- more at bide
Date: 13th century
1 a : allegiance to duty or a person : loyaltyb(1) : fidelity to one's promises(2) : sincerity of intentions
2 a(1) : belief and trust in and loyalty to God(2) : belief in the traditional doctrines of a religionb(1) : firm belief in something for which there is no proof(2) : complete trust
3 : something that is believed especially with strong conviction; especially : asystem of religious beliefs
-in faith : without doubt or question : verily
1
Synonyms BELIEF 1, credence, credit
Contrasted Words dubiety, dubiosity, skepticism, uncertainty
2
Synonyms TRUST 1, confidence, dependence, hope, reliance, stock
Contrasted Words disbelief, incredulity, unbelief; apprehension, misgiving
3
Synonyms RELIGION 1, creed, cult, persuasion
4
Synonyms RELIGION 2, church, communion, connection, creed, cult, denomination, persuasion, sect
Related Word doctrines, dogmas, tenets

With one, that in her sex, her years, profession,
Wisdom, and constancy, hath amaz'd me more
Than I dare blame my weakness.
[William Shakespeare: All's Well That Ends Well, Ñ. 40]

Cette haute fréquence de l'emploi de *faith* et son haut degré de la désémantisation dans les groupes de mots prépositionnels, témoignent du caractère organique de ce lexème pour tout le corpus du vocabulaire des XVI-XVII siècles.
Le deuxième groupe (115 cas) concerne des textes exprimant les *relations de contrat*: contrat d'affaire avec un partenaire, contrat entre époux (ou amoureux), contrat entre les hommes et Dieu, contrat entre un seigneur féodal et ses vassaux. Constatons le fait dans les exemples suivants :

1*. The bargain of your faith, I do beseech you*
Even at that time I may be married too.
[William Shakespeare: The Merchant of Venice, Ñ. 88]
2. *PRO.*
Already have I been false to Valentine,
And now I must be as unjust to Thurio:
Under the color of commending him,
I have access my own love to prefer -
But Silvia is too fair, too true, too holy,
To be corrupted with my worthless gifts.
When I protest true loyalty to her,
She twits me with my falsehood to my friend;
When to her beauty I commend my vows,
She bids me think how I have been forsworn
In breaking faith with Julia whom I lov'd;
And notwithstanding all her sudden quips,
The least whereof would quell a lover's hope,
Yet, spaniel-like, the more she spurns my love,
The more it grows, and fawneth on her still.
[William Shakespeare: The Two Gentlemen of Verona, Ñ. 104]
3. In him that was of late an heretic,
As firm as faith.
[William Shakespeare: The Merry Wives of Windsor, Ñ. 127]
3a.K. RICH.
They break their faith to God as well as us.
[William Shakespeare: The Tragedy of King Richard the Second, Ñ. 82]

C'est à la réalisation de la signification à l'intérieur des collocations par excellence qu'il faut prêter attention pendant la présentation des contextes du lexème *faith* dans le sens du contrat[1] . Sur 115 cas, 87 sont soit en combinaison avec un verbe (le plus souvent, dans la forme impersonnelle, exprimée par un participe passé), soit avec un adjectif qualificatif. Signalons un cas de l'emploi avec un substantif en génitif (*virgin 's faith).*

[1] Collocation f, la contextualisation au niveau lexique, c'est un entourage lexique typique.

collocations verbales	**collocations adjectivales, nominales et participiales**
to awake faith *to break faith* *to call faith* *to cast faith* *to defy faith* *to descry faith* *to discard faith* *to disjoin faith* *to disparage faith* *to doubt one's faith* *to fall from one's faith* *to fix faith with smth* *to forfeit / unforfeat faith* *to forswear faith* *to give faith* *to hold faith* *to infringe faith* *to lose faith* *to plight faith* *to praise faith* *to put from faith* *to seal faith* *to swear / unswear faith* *to tear faith* *to trod faith down* *to vow faith* *to wear faith*	*besmear'd and over-stain'd faith ** *broken faith** *dear faith* *deep-sworn faith** *discarded faith*.* *fair faith* *false faith* *firm faith* *honest faith* *irrevocable faith* *little faith* *plain and simple faith* *plighted faith** *plural faith* *a saving faith* *strong faith* *true faith* virgin 's faith (n)

Tableau 1. Collocations verbales et adjectivales du substantif faith dans les textes de Shakespeare

Au total, nous avons relevé dans les textes de Shakespeare 27 collocations verbales, 12 collocations adjectivales et 1 collocation nominale. Les collocations verbales forment 5 groupes: (*1) to give faith (to swear faith) (2) to break faith (forswear, infringe,) (3) to hold faith (to plight faith, to vow faith, to fix faith) (4)to lose (disparage, to fall from faith, to disjoin faith, to forfeit faith, to put from faith, to discard faith), (5)to challenge faith (defy).* Les collocations adjectivales réalisent l'opposition suivante: *true (fair, honest, strong, etc.) faith / false faith (little, besmeared)*[1] .Citons quelques contextes des oeuvres de Shakespeare:

[1] Cf. les combinaisons du lexème *faith*, fixées dans le dictionnaire anglais combinatoire (M. Bensona, E. Benson, R. Ilson, 1990): 1.to have faith in, 2. to place one's faith in, 3.to lose faith in, to shake smb.'s faith in, 4/ an abiding, enduring, steadfast, deep, strong, unshakable faith, on faith to accept on faith, to keep faith with, to demonstrate, show good faith, in good faith, in bad faith, to adhere, practice a faith, to abjure, recant, renounce one's faith, the true faith, by faith.

collocations verbales	collocations adjectivales
1."If love make me forsworn, how shall I swear to love? *Ah, never faith could hold, if not to beauty vowed!* *Though to myself forsworn, to thee I'll faithful prove;* *[William Shakespeare: Love's Labor's Lost, Ñ. 76]* *2.PRIN.* *And quick Berowne hath plighted faith to me.* *[William Shakespeare: Love's Labor's Lost, Ñ. 132]* *3.DEM.* *Disparage not the faith thou dost not know,* *Lest, to thy peril, thou aby it dear.* *Look where thy love comes; yonder is thy dear.* *[William Shakespeare: A Midsummer Night's Dream, Ñ. 68]*	*1.Than plural faith, which is too much by one.* *Thou counterfeit to thy true friend!* *[William Shakespeare: The Two Gentlemen of Verona, Ñ. 140]* *2.Few words to fair faith. Troilus shall be such to Cressid as what envy can say worst shall be a mock for his truth, and what truth can speak truest not truer than Troilus.* *CRES. Will you walk in, my lord?* *TRO. You know now your hostages: your uncle's word and my firm faith.* *[William Shakespeare: The History of Troilus and Cressida, Ñ. 96]* *3.There are no tricks in plain and simple faith;* *But hollow men, like horses hot at hand,* *Make gallant show and promise of their mettle;* *[William Shakespeare: The Tragedy of Julius Caesar, Ñ. 113]*

Tableau 1 bis. Collocations verbales et adjectivales du substantif faith dans les textes de Shakespeare (exemples de contextes)

La sémantique et les collocations du lexème *faith* indiquent certains changements à l'intérieur du concept de base qu'il réalise. On peut suivre ces changement en se référant avant tout à l'étymologie du mot *faith* et à celle des verbes des collocations. Ce mot, d'après l'encyclopédie *Britannica Deluxe*, s'est fixé en anglo-normand au XIII-e siècle. Il a pénétré le vocabulaire de l'ancien français après la Conquête de l'Angleterre par les Normands. Il vient du mot latin *fides*[1] qui signifiait une des valeurs principales de la société romaine, mythologisée sous forme d'une divinité abstraite, allégorisée comme la déeese du cercle de Jupiter qui répond de l'intégrité morale de la société romaine. La réalisation de relations contractuelles entre les dieux et les hommes était l'une des fonctions principales de cette déesse. A la période postérieure du développement de la société romaine on attribua à cette déesse la fonction de la surveillance des contrats et des documents[2]. La signification «foi» (appartenance à une confession religieuse) qui domine dans les dictionnaires et les textes contemporains, n'apparaît qu'avec le christianisme.

Comme nous venons de le voir, le «sème des relations contractuelles» est le sème principal dans l'emploi du lexème *faith* dans les oeuvres de Shakespeare. En fait, son emploi est proche de l'emploi terminologique. Un tel emploi apparaît clairement dans les «Récits de Kenterbery» de

[1] Etymology: Middle English feith, from Old French feid, foi, from Latin *fides*; akin to Latin fidere to trust- more at bide
Date: 13th century.

[2] FIDES
Roman goddess, the deification of good faith and honesty. Many of the oldest Roman deities were embodiments of high ideals (e.g., Honos, Libertas); it was the function of Fides to oversee the moral integrity of the Romans. Closely associated with Jupiter, Fides was honoured with a temple built near his on the Capitoline Hill in 254 BC. In symbolic recognition of the secret, inviolable trust between gods and mortals, attendants presented sacrificial offerings to her with covered hands.

In the later Roman period, she was called Fides Publica ("Public Faith") and was considered the guardian of treaties and other state documents, which were placed for safekeeping in her temple. There, too, the Senate often convened, signifying her importance to the state. (Britannica Deluxe 2004).

Chaucer (5 cas de 30)[3]. Citons l'exemple le plus caractéristique où le sème du contrat entre un homme et une femme (fidélité) se réalise :

I have thy faith and thy benyngnytee
As wel as evere womman was, assayed
In greet estaat, and povreliche arrayed;
Now knowe I, goode wyf, thy stedfastnesse!"
And hir in armes took, and gan hir kesse.
Signalons que chez Chaucer le sème «foi en Dieu» se réalise le plus souvent (15 cas de 30)[1] :
For hooly chirches faith in oure bileve.
Une telle réalisation est liée, premièrement, à la structure du héros (un chevalier) et, deuxièmement, aux particularités épistémiques de la culture médiévale, quand les relations d'affaire, les relations commerciales entre des partenaires n'occupent pas encore le premier plan.
10 cas de l'emploi d'un groupe de mots prépositionnel *("In faith, Squier, thow hast thee wel yquit,And gentilly I preise wel thy wit")* témoignent que déjà vers le XIVe siècle le lexème *faith* s'est si bien fixé dans la langue qu'il commence à figurer dans des groupes de mots désémantisés.

Une question se pose : si le lexème *faith,* qui avait pénétré l'anglais au XIII-e siècle et s'était inscrit organiquement dans le vocabulaire, exprimait les relations contractuelles, quel lexème avait-t-il remplacé? Il est peu probable qu'il ait rempli une lacune conceptuelle, puisque les relations contractuelles sont les premières relations qui s'établissent dans une société, et se réflètent dans les systèmes mytho-poétiques (voir G. Frazer «Folklore dans le Vieux Testament», M.I. Stebline-Kamenskij «Le mythe et le devenir de la littérature», E. Meletinskij[2]). Elles caractérisent toutes les cultures et il est peu probable que sur le territoire de l'Angleterre, avant la Conquête par les Normands, il n'y ait pas eu d'institutions réglant ces relations, peu probable également qu'en anglais il ne reste pas des traces de l'expression des vieux régulateurs. En quête de ces traces, référons-nous à l'étymologie des verbes qui font partie des collocations de Shakespeare avec le lexème *faith*, en excluant les verbes connotés affectivement.

	Verbe	**Etymologie. La date de l'entrée**
1*.1.*	*to break faith*	*Etymology: Middle English breken, from Old English brecan; akin to Old High German brehhan to break, Latin frangere* *Date: before 12th century*
22.	*to call faith*	*Etymology: Middle English, from Old Norse kalla; akin to Old English hildecalla battle herald, Old High German kallʓn to talk loudly, Old Church Slavonic glas V voice* *Date: before 12th century*
33.	*to cast faith*	Etymology: Middle English, from Old Norse kasta; akin to Old Norse kqs heap Date: 13th century
44.	*to defy faith*	*Etymology: Middle English, to renounce faith in, challenge, from Middle French defier, from de- + fier to entrust, from (assumed) Vulgar Latin fidare, alteration of Latin fidere to trust— more at bide* *Date: 14th century*
55.	*to descry faith*	*Etymology: Middle English descrien, from Middle French descrier to proclaim, decry* *Date: 14th century*

[3] En tout 30: 10 – la religion, 15 – groupes de mots prépositionnels, 5 – les relations contractuelles. La révision a été réalisée sur la base de la bibliothèque électronique «Trésors de la littérature mondiale» qui contient un programme de recherche.
[1] Signalons que dans les tragédies de Christopher Marlowe 28 cas de 32 cas de l'emploi du lexème *faith* réalisent la signification «religion», «foi en Dieu».
[2] Meletinskij E.M. Iwbrannyie statji. Vospominanija. (Articles choisis. Souvenirs). Moskva: Izdatelstvo Rossijskogo gumanitarnogo universiteta, 1998; Stebline-Kamenskij M.I. Mir sagi. Stanovlenije literatury (Le monde de la saga. Le devenir de la littérature). Leningrad: Nauka, 1984; Frazer G.G. Folklor v Vetkhom zavete (Folklore dans le Vieux Testament). Moskva: Politizdat, 1989.

66.	*to discard faith*	*Etymology: Middle English carde, from Middle French, from Late Latin cardus thistle, from Latin carduus— more at chard* *Date: circa 1586*
77.	*to disparage faith*	*Etymology: Middle English, to degrade by marriage below one's class, disparage, from Middle French desparagier to marry below one's class, from Old French, from des- dis- + parage extraction, lineage, from per peer* *Date: 14th century*
88.	*to fall from one's faith*	*Etymology: Middle English, from Old English feallan; akin to Old High German fallan to fall and perhaps to Lithuanian pulti* *Date: before 12th century*
99.	*to fix faith with smth*	*Etymology: Middle English, from Latin fixus, past participle of figere to fasten; akin to Lithuanian dygti to sprout, break through* *Date: 14th century*
110.	*to forfeit / unforfeat faith*	*Etymology: Middle English forfait, from Middle French, from past participle of forfaire to commit a crime, forfeit, from fors outside (from Latin foris) + faire to do, from Latin facere— more at forum, do* *Date: 14th century*
111.	*to forswear faith*	*Etymology: Middle English forsweren, from Old English forswerian, from for-+ swerian to swear* *Date: before 12th century*
112.	*to give faith*	*Etymology: Middle English, of Scandinavian origin; akin to Old Swedish giva to give; akin to Old English giefan, gifan to give, and perhaps to Latin hab‡re to have, hold* *Date: 13th century*
113.	*to hold faith*	*Etymology: Middle English, from Old English healdan; akin to Old High German haltan to hold, and perhaps to Latin celer rapid, Greek klonos agitation* *Date: before 12th century*
114.	*to infringe faith*	*Etymology: Medieval Latin infringere, from Latin, to break, crush, from in- + frangere to break— more at break* *Date: 1533*
115.	*to lose faith*	*Etymology: Middle English, from Old English losian to perish, lose, from los destruction; akin to Old English l‡osan to lose; akin to Old Norse losa to loosen, Latin luere to atone for, Greek lyein to loosen, dissolve, destroy* *Date: before 12th century*
116.	*to plight faith*	*Etymology: Middle English, from Old English plihtan to endanger, from pliht danger; akin to Old English pl‡on to expose to danger, Old High German pflegan to take care of* *Date: 13th century*
117.	*to put from faith*	*Etymology: Middle English putten; akin to Old English putung instigation, Middle Dutch poten to plant* *Date: 12th century*
118.	*to seal faith*	*Etymology: Middle English sele, from Old English seolh; akin to Old High German selah seal* *Date: before 12th century*
119.	*to swear / unswear faith*	*Etymology: Middle English sweren, from Old English swerian; akin to Old High German swerien to swear and perhaps to Old Church Slavonic svarV quarrel* *Date: before 12th century*
220	*to vow faith*	*Etymology: Middle English vowe, from Old French vou, from Latin*

.		*votum, from neuter of votus, past participle of vov‡re to vow; akin to Greek euchesthai to pray, vow, Sanskrit v†ghat sacrificer Date: 14th century*

Tableau 2

L'analyse étymologique a montré que seulement huit (4, 5, 6, 7, 9, 10, 14, 20) des vingt verbes, qui font partie des collocations, viennent du latin à travers l'ancien français, et sont apparus dans la langue anglaise après la Conquête de l'Angleterre par les Normands. Les douze autres verbes sont d'origine germano-scandinave et se fixèrent dans la langue avant le XIIe siècle. Une telle distribution reflète la situation juridique réelle qui existait sur le territoire de l'Angleterre dans le haut Moyen Age: c'est-à-dire la distribution en territoires sous la juridiction de la loi anglo-saxone (Anglo-Saxon Law) et en territoires sous la juridiction de la loi danoise (Dane Law). Dans les lois du roi Alfred (871-901 A.D.) qui gouverne le territoire de la loi anglo-saxone et dans le traité connu comme «**The Laws of Alfred, Guthrum, and Edward the Elder**», où les relations entre les deux territoires[1] sont réglées, nous trouvons le terme ***wжr***, utilisé, comme le montre le dictionnaire anglo-saxon, pour désigner les relations contractuelles (e; f – covenant, compact, agreement, pledge)[2]. Outre cela, ce terme, comme les exemples des textes de manuscrits donnés par l'article du dictionnaire le montrent, avait désigné les relations contractuelles dans les premières traductions de textes bibliques en anglo-saxon.

Cp. *Wær is ætsomne Godes and monna, gestæhālig treōw, Exon.Th* **[**Bosworth J. An ANGLO-SAXON DICTIONARY - 1156].

Ce mot, d'après le dictionnaire historique et étymologique du russe moderne de Tchernykh P.J. (P. 141), remonte, à son tour, à la racine indo-européenne ∗vera, ∗ uērā et se trouve pratiquement dans toutes les langues indo-européennes modernes. En russe cette racine se retrouve dans les mots *vera, doverije, doverennost* et renvoie aux mêmes significations que le mot moderne *faith*[3] . En latin ce radical s'est retrouvé dans le lexème *vērus,* bien que les relations contractuelles fussent liées au mot *fides*, qui donna «foi» en français et *faith* en anglais. Le mot anglo-saxon ***wжr*** s'est conservé seulement dans l'homonyme *Warwick*. On peut expliquer la substitution complète du référent (le signifié) pour le signifiant , premièrement, par le changement de la langue des auteurs de la loi. Le roi Alfred, et le roi Edward rédigeaient des lois en anglo-saxon. Après la Conquête de l'Angleterre par les Normands les rois rédigèrent leurs lois soit en ancien français, soit en latin, soit en anglo-normand. Deuxièmement, si les lois anglo-saxones reflétaient foncièrement les relations féodales avec des vestiges de l'esclavage, après le XIIIe siècle les lois royales commencèrent à refléter plus les relations d'affaire et de mariage, ce qu'on voit surtout quand on analyse l'emploi du lexème *faith* dans les textes de Shakespeare. En tout cas, le lexème ***wжr*** avait disparu avec ses utilisateurs. La comparaison de l'emploi de ce lexème dans les dictionnaires montre que le vecteur de la signification «foi en DIEU, religion» après la Renaissance, croissait, tandis que le vecteur de la signification «relations de partenaire» baissait. Tout cela exige une analyse de l'usage du lexème *faith* dans les textes des XVII-XIX siècles, ce qui est au-delà de notre recherche, mais peut devenir l'objet d'un autre ouvrage.

Pour en revenir au thème de notre recherche, tel que nous l'avons présenté au début de cet exposé : – définir quelles significations Shakespeare avait mises dans le mot *faith* dans la ligne «*And purest faith unhappily forsworn»*, signalons la possibilité de la réalisation de toute la sémantique du mot, ce qui caractérise en principe la langue de Shakespeare. Pourtant la collocation verbale du mot *faith* (*forsworn*) lie cette ligne à trois autres textes, où ce lexème réalise le sème des relations de partenaire entre les amoureux :

---

[1] These are the dooms which King Alfred and King Guthrum chose. And this is the ordinance also which King Alfred and King Guthrum, and afterwards King Edward and King Guthrum, chose and ordained, when the English and Danes fully took to peace and to friendship; and the witan also, who were afterwards, oft and unseldom that same renewed and increased with good. [**The First Written Laws of The Anglo-Saxons URL: http://www.ealdriht.org,** Maitland, F.W. "The Laws of the Anglo-Saxons," The Collected Papers of Frederic William Maitland, ed. by H.A.L. Fisher, vol. 3, pp. 447–473 (1911, reprinted 1981].

[2] Bosworth J. An ANGLO-SAXON DICTIONARY based on the manuscript collections. Oxford Univesity Press,- 1996, pp.1156-1158.

[3] Cf. le dernier verset du 13 chapitre de la Première épître aux Corinthiens: Maintenant donc demeurent foi, espérence, charité/I nyne prebyvaiut Vera, Nadejda, Liubov/ and now abideth Faith, Hope and Charity (Corinth. 13.13).

1."If love make me forsworn, how shall I swear to love? *Ah, never faith could hold, if not to beauty vowed!* *Though to myself forsworn, to thee I'll faithful prove;* *[William Shakespeare: Love's Labor's Lost, Ñ. 76]* 3.. *If love make me forsworn, how shall I swear to love?* *O, never faith could hold, if not to beauty vowed:* *Though to myself forsworn, to thee I'll constant prove;* [William Shakespeare: The Passionate Pilgrim, Ñ. 6]	*2. In loving thee thou know'st I am forsworn,* *But thou art twice forsworn, to me love swearing;* *In act thy bed-vow broke, and new faith torn* *In vowing new hate after new love bearing.* *But why of two oaths' breach do I accuse thee,* *When I break twenty? I am perjur'd most,* *For all my vows are oaths but to misuse thee,* *And all my honest faith in thee is lost;* *For I have sworn deep oaths of thy deep kindness,* *Oaths of thy love, thy truth, thy constancy,* *And to enlighten thee gave eyes to blindness,* *Or made them swear against the thing they see;* *For I have sworn thee fair: more perjur'd eye,* *To swear against the truth so foul a lie!* *[William Shakespeare: Sonnets, Ñ. 153]*

On peut considérer ces textes comme les textes-clés pour le décèlement de la sémantique du lexème *faith* dans la ligne analysée. Avec cela, signalons que le vers de la comédie «Les efforts vains de l'amour» (1) est complètement identique, sauf la paire variable *faithful/ constant* dans la troisième ligne, aux lignes du «Pèlerin passionné» (3). Le cent-cinquante-troisième sonnet (2) où il s'agit d'amour et de trahison, réalise tout le faisceau du thésaurus des relations contractuelles d'amour avec l'emploi de presque toutes les collocations signalées.

Ainsi, sans sortir des limites de l'analyse déconstructive, c'est-à-dire en reconstituant les significations à l'aide de l'analyse étymologique, définitionnelle, de constituants et contextuelle, et en considérant la ligne du 66ème sonnet de Shakespeare dans sa corrélation avec toutes ses oeuvres et en sélectionnant pour la comparaison les oeuvres de Chaucer et de Christopher Marlowe (analyse intertextuelle), nous avons réussi à définir la stabilité du concept «relations contractuelles», qui pouvait se réaliser en lexèmes différents, et la sémantique du lexème *faith*, conditionnée épistémologiquement, dans les textes de Shakespeare, ce qui témoigne de l'homogénéité de la paternité des textes. Avec cela, les méthodes de la linguistique de corpus (utilisation de 4 bibliothèques électroniques sur les CD et aussi de sources de l'Internet, de programmes de recherche) ont aidé à suivre le développement de la notion en général et à définir les vecteurs du développement retrospectivement et prospectivement. A notre avis, l'application des méthodes de la linguistique de corpus a assuré l'authenticité de l'attribution de la paternité de ses œuvres à Shakespeare.

ÉTUDE QUANTITATIVE DES CHANGEMENTS ESTHÉTIQUES ET DES VARIATIONS GÉNÉRIQUES CHEZ TROIS GRANDS ÉCRIVAINS : ANALYSE LEXICOMÉTRIQUE D'UN CORPUS LITTÉRAIRE

Margareta KASTBERG SJÖBLOM
ILF-CNRS, BCL, Université de Nice

SOMMAIRE

La notion de genre reste encore aujourd'hui l'institution première du code littéraire, bien qu'elle ait souvent été discutée et remise en question. Les théoriciens la considèrent avec réserve, affirmant que chaque genre littéraire en englobe plusieurs, et les hésitations terminologiques manifestent ce caractère "d'appartenance multiple et emboîtante" de tout écrit littéraire. En effet, la codification des genres n'est pas chose aisée ni stabilisée. Le système traditionnel nous propose – ou nous impose – selon le code générique institutionnel, certaines classifications reconnues : romans, nouvelles, essais, etc. Mais, cette distinction des genres transmise par la critique littéraire "traditionnelle" est-elle réellement pertinente ?
Pour répondre à cette question essentielle, nous avons choisi de recourir à l'outil informatique et aux méthodes de la linguistique quantitative qui montrent bien que les genres existent, qu'on le veuille ou non, et qu'il serait inconcevable sur le plan purement linguistique de nier l'existence de différentes typologies de textes. L'analyse lexicométrique valide cette idée, l'opposition générique est extrêmement claire et permet de définir des caractéristiques génériques en s'appuyant, non sur des valeurs anthropologiques ou sociales, mais sur les propriétés mêmes des textes.
Le présent exposé propose d'étudier les variations et les oppositions génériques chez plusieurs grands écrivains français : Julien Gracq, Gustave Flaubert et J.M.G. Le Clézio, et en s'appuyant sur un corpus informatisé et lemmatisé, et en exploitant les techniques quantitatives. L'œuvre de ces écrivains présente une riche variété de textes et se décline en différents genres, ayant des caractéristiques typologiques bien distinctes. Bien que les auteurs évoquent souvent une "écriture unique", déclarent n'appartenir à aucun groupe et tentent même de transgresser un système social établi, les différentes typologies de textes existent et ces variations sont à observer à tous les niveaux.
L'analyse du corpus en situation permet en effet d'abord de caractériser la structure du vocabulaire. Le rythme de la narration est ensuite corrélé à l'analyse de la longueur des mots et des phrases. Enfin, l'étude de la distance lexicale met en évidence des aspects sémantiques et thématiques révélateurs et permet aussi de constater d'importantes variations typologiques. Ces différentes analyses mettent donc bien en exergue l'opposition entre les différents genres, toujours présente et souvent même prépondérante dans toutes les différentes analyses statistiques.

1. Le corpus

Plusieurs écrivains français ont mis en question ou refusent même le cloisonnement en genres, parlant d'une seule et unique écriture. Parmi ces auteurs certains ont une large production qui se décline en plusieurs genres littéraires. Nous nous intéresserons ici à trois d'entre eux : Julien Gracq, Jean-Marie Le Clézio et Gustave Flaubert.
Un point commun entre ces trois écrivains très productifs est la difficulté de classer leurs ouvrages dans des genres littéraires traditionnels. Notamment la critique gracquienne évite parfois complètement de prendre position ou d'effectuer une classification générique quelconque en qualifiant tous les textes de palimpsestes.
Notre corpus Gracq, englobe pratiquement la totalité de sa production avec 17 ouvrages et rassemble plusieurs genres littéraires, notamment les essais littéraires qui sont richement représentés dans ce corpus :

Romans : *Au château d'Argol*, *Un beau ténébreux*, *Le rivage des Syrtes*, *Un balcon en forêt* et *La presqu'île*.
Critiques, Essais ou mélanges[1] : *André Breton. Quelques aspects de l'écrivain*, *Préférences*, *Lettrines 1*, *Lettrines 2*, *Les eaux étroites, En lisant, en écrivant*, *La forme d'une ville*, *Autour des sept collines* et *Carnets du grand chemin*.
Poèmes en prose : *Liberté grande*.
Théâtre : *Le Roi-pêcheur* et *Penthésilée*.

La production littéraire de Le Clézio est vaste, s'étend sur plus de quarante ans et englobe plusieurs genres littéraires. Il est constitué de 31 livres ; tout d'abord des six premières œuvres, classées, par leur style particulier et innovant, comme appartenant à l'École du "nouveau roman" : *Le procès-verbal*, *La fièvre*, *Le déluge*, *Le livre des fuites*, *La guerre* et *Voyages de l'autre côté*. Les neuf romans qui suivent cette période, considérés par les critiques comme plus "traditionnels", sont les suivants : *Désert*, *Le chercheur d'or*, *Voyage à Rodrigues* (écrit sous forme de journal personnel), *Angoli Mala*, *Onitsha*, *Etoile errante*, *La quarantaine*, *Poisson d'or* et *Hasard*. *Mydriase* et *Vers les icebergs* sont difficiles à classer dans un genre précis, ce sont plutôt des récits poétiques. Le corpus inclut ensuite les recueils de nouvelles : *Mondo et autres histoires*, *La ronde et autres faits divers* et *Printemps et autres saisons*. Les essais littéraires sont de différentes époques. *L'extase matérielle* et *L'inconnu sur la terre* traitent de thèmes généraux tandis que *Trois villes saintes* et *Le rêve mexicain ou la pensée interrompue* s'intéressent exclusivement à la culture amérindienne. Celle-ci constitue également le principal intérêt des ouvrages à vocation ethnologique, *Les prophéties du Chilam Balam* et *La fête chantée*, tandis que *Sirandanes* s'intéresse à la culture de l'île Maurice. Sont inclus en outre dans le corpus deux livres pour enfants : *Voyage au pays des arbres* et *Pawana* ; la seule biographie *Diego et Frida*, et le récit de voyage *Gens des nuages*.
La production de Gustave Flaubert s'étend sur la moitié du XIXème siècle et ce corpus inclut 15 livres ; des romans comme *Madame Bovary*, *L'éducation sentimentale*, *Salammbô* ou *Bouvard et Pécuchet*. Nous y trouvons aussi *Les trois contes*, le dialogue poétique et philosophique de *La tentation de Saint Antoine* dans ses trois versions, le récit de voyage *Par les champs et par les grèves,* les *Mémoires* et les *Souvenirs* ainsi qu'une partie de la *Correspondance*.
Ce grand corpus a été numérisé et traité par le logiciel *Hyperbase*, version 6.0. et il contient 4.121.141 occurrences et 79.833 formes (dans la version qui s'appuie sur les formes graphiques) réparties sur les soixante-trois œuvres du corpus.
Le traitement lexicostatistique automatisé permet un certain nombre d'analyses qui ouvrent la voie à des interprétations et à des études différentes de ce corpus, basées sur des données impartiales, et non sur des critères subjectifs.
C'est en premier lieu à travers une étude sur la structure lexicale du corpus que nous pouvons observer l'influence de la riche variation typologique des textes.

2. La structure lexicale

Les différentes recherches sur la structure lexicale offrent la possibilité, indépendamment du contenu lexical, de situer, de distinguer et comprendre la structure formelle des textes afin de pouvoir comparer différents discours, genres, époques ou auteurs, au niveau exogène aussi bien qu'au niveau endogène, ainsi que les parties de l'œuvre d'un écrivain ou de tout autre producteur de texte ou de parole. Ces recherches, qui au fond sont très proches de la lexicométrie traditionnelle, permettent aussi d'étudier l'évolution dans le temps.
Les calculs effectués par le logiciel *Hyperbase*, utilisé dans cette étude, permettent de mesurer l'étendue des textes dans le corpus en prenant en compte des contraintes statistiques. Les calculs du poids relatif, c'est-à-dire l'espérance mathématique de l'événement : occurrence d'un mot dans le texte considéré (P) et non-occurrence de ce mot dans le même texte (Q=1-P), permettent l'emploi des lois classiques de la lexicométrie, principalement la loi normale et la loi binomiale (Muller 1977 : 159-169), et elles servent aux calculs de pondération dans les différents traitements statistiques.
L'étude de l'accroissement lexical détermine l'apport du vocabulaire au fil du temps ; cet accroissement est, pour un segment déterminé du texte, le nombre d'unités nouvelles, c'est-à-dire n'ayant pas été employées antérieurement, qui apparaissent dans ce segment. Pour effectuer

[1] Gracq donne souvent la dénomination de "fragments" à ces ouvrages.

cette mesure, on découpe le corpus en tranches. La représentation graphique ci-dessous rend compte de l'accroissement du vocabulaire dans l'ordre chronologique.

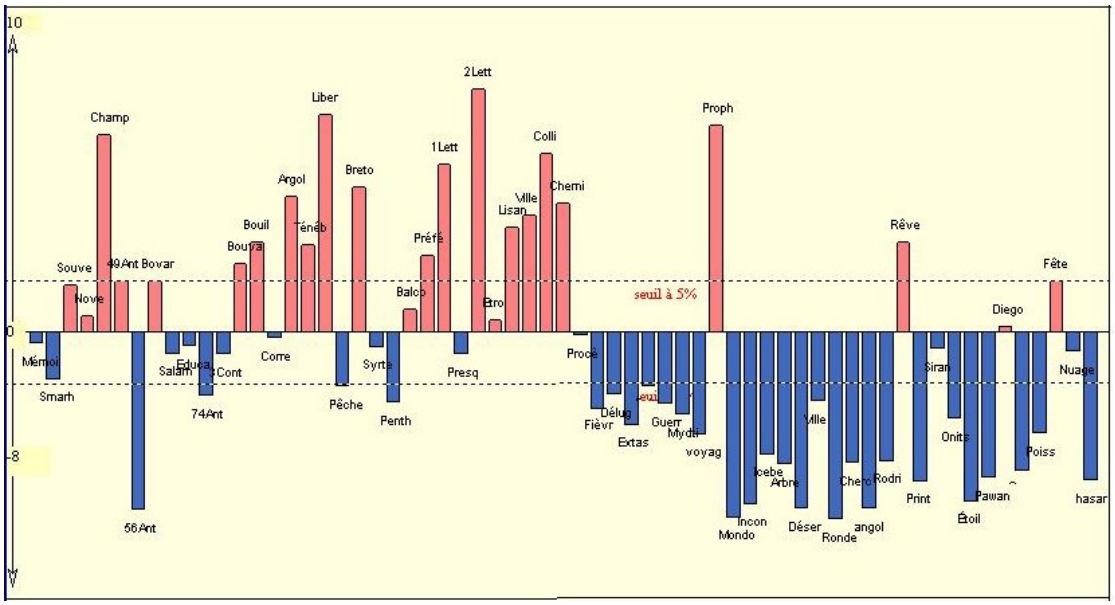


Figure n°1 : Accroissement lexical du corpus

Le graphique qui, de gauche à droite, s'oriente selon la chronologie, avec le corpus Flaubert suivi par celui de Gracq et de Le Clézio[1] nous permet de constater que les écarts autour de la moyenne, l'axe horizontal, sont de très grande ampleur, avec des ruptures et des reprises. Le seuil à 5 % est dépassé de nombreuses fois, avec des "pics" importants, dans le sens positif aussi bien que négatif.

L'étude de l'accroissement fait en effet très clairement apparaître l'opposition générique très importante du corpus. Dans la première partie, chez Flaubert, le vocabulaire ne croît de façon significative qu'une seule fois avec l'unique récit de voyage *Par les champs et par les grèves.*

Parmi les ouvrages de Gracq, notons que le récit *Liberté grande* et les essais comme *Lettrines I et II* ainsi que *Autour des sept collines* et *Carnets du grand chemin*, introduisent régulièrement de nouveaux thèmes dans le corpus. Mais le plus frappant est peut-être l'extraordinaire impact de l'apport lexical qui advient avec l'introduction du monde amérindien dans le corpus, ici avec l'essai littéraire *Le rêve mexicain* de Le Clézio. Notons aussi l'introduction d'un nouveau genre à ce corpus ; celui de l'ouvrage ethnologique qui fait appel à un apport lexical massif dans *Les prophéties de Chilam Balam*.

Nous pouvons en effet constater que dans ce corpus l'opposition des genres est plus importante que celle des trois auteurs. Il n'y a pas de limite nette entre l'œuvre de Flaubert et celle de Gracq. En revanche, la partie qui couvre les ouvrages de Gracq est nettement plus riche en apport de vocabulaire que celle de Le Clézio.

Par ailleurs, l'étude de la richesse lexicale[2] a montré que Gracq semble avoir un usage plus riche du vocabulaire tandis que Le Clézio s'exprime en règle générale avec un vocabulaire plus restreint. C'est dans la partie gracquienne que nous constatons des valeurs excédentaires et plus précisément dans la dernière partie et dans les essais, les pièces de théâtre étant forcément pauvres, qui témoignent d'un vocabulaire dont la richesse augmente vers la fin de l'œuvre.

3. Le rythme du récit

La ponctuation est essentiellement d'ordre syntaxique et doit être comprise comme l'écrit Nina Catach[3] :

> "… associant à la fois : une suite de mots (aspect constructif), un message (aspect actuel), une substance et une forme intonatives (mélodie expressive et aspect intonatif) et un sens

[1] Il convient, avant d'interpréter cet histogramme, de souligner le fait que ce corpus n'est pas chronologique, et en postposant Le Clézio à Gracq cette étude désavantage évidemment le dernier auteur.

[2] Cf. M. Kastberg Sjöblom (2006) p. 50-57.

[3] N. Catach (1994) p. 48.

(contenu du message, résultant de l'ensemble des données précédentes). La ponctuation tient dignement son rôle à ces différents niveaux de la syntaxe."

Dans le langage écrit, la fin de la phrase est généralement représentée par le point, les points de suspension, le point d'interrogation, le point d'exclamation.
L'emploi des signes de ponctuation varie selon le genre de discours, l'époque et l'auteur. La ponctuation peut également connaître de grandes variations chez un même écrivain. L'étude de la distribution de signes de ponctuation permet de définir la longueur de la phrase et ainsi le rythme du récit. Le logiciel *Hyperbase* permet de recenser les différents signes de ponctuation de manière aisée et fiable et permet ainsi de connaître la longueur de la phrase. La longueur de la phrase et les caractéristiques de ses segments dépendent étroitement de sa complexité. Le recensement des signes de la ponctuation faible permet de connaître la segmentation interne de la phrase et d'étudier sa complexité. Toutefois, dans cette étude nous nous bornons au recensement des signes de ponctuation forte. Cette distribution n'est pas régulière à travers le corpus, les bâtons excédentaires en signes de ponctuation révèlent donc une phrase courte et brève tandis que les bâtons bleus, déficitaires témoignent des phrases longues.

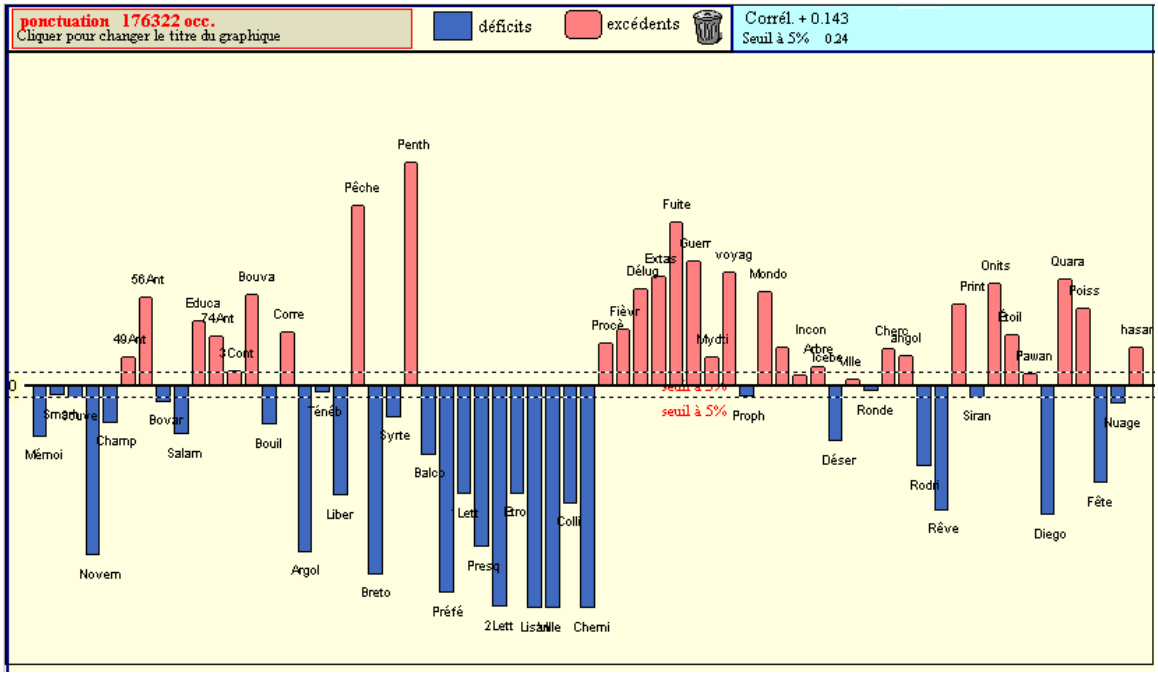


Figure n° 2 : Distribution relative des quatre signes de ponctuation forte dans le corpus

L'histogramme distingue bien les différences entre les auteurs : Flaubert semble abandonner une phrase longue et complexe pour des phrases plus courtes, tandis que le récit gracquien est plus généralement caractérisé par une phrase longue. Quant à l'écriture leclézienne on distingue les trois parties caractéristiques son œuvre, si souvent commentées par les critiques.
Néanmoins, c'est peut-être le facteur du genre qui est le plus révélateur de cet histogramme. Le début romanesque de Flaubert est caractérisé par des phrases longues, tandis que des ouvrages comme *La tentation de Saint Antoine* font appel à une phrase plus courte, tout comme les romans tardifs et la correspondance. La phrase de Gracq est longue, à l'exception des pièces de théâtre *Le Roi-pêcheur* et *Penthésilée*. Avec Le Clézio on introduit à ce corpus le genre particulier du nouveau roman qui est caractérisé par une phrase courte et brève. Lorsque Le Clézio abandonne ce genre pour un style romanesque plus "traditionnel" la phrase reste généralement courte. C'est d'ailleurs une des caractéristiques du style leclézien, apprécié intuitivement par les lecteurs. C'est dans les ouvrages ethnologiques, les essais littéraires et la biographie *Diego et Frida* que nous trouvons des phrases longues correspondant à un style bien plus intellectuel.
En effet, nous avons pu observer, dans les diverses analyses sur le rythme de la phrase que Le Clézio exploite différentes techniques et différents styles d'expression selon les époques et selon les différents genres littéraires et que les résultats qui en découlent manifestent des variations non négligeables.

Le facteur prédominant de ces divergences semble en effet être celui du genre. Les mots courts dominent l'œuvre romanesque tandis que les mots longs se trouvent dans les ouvrages ethnologiques et dans les essais. C'est également à l'intérieur de ce genre que nous trouvons les phrases les plus longues qui toutefois ne manifestent pas de complexité particulière, préférant la coordination à la subordination[1].
Les analyses que nous avons effectuées jusqu'à présent ont en commun de ne pas considérer le mot en soi, mais des liens statistiques qui donnent à voir des réseaux signifiants indépendants de l'interprétation du contenu. Dorénavant, nous nous intéresserons au contenu du discours qui implique la signification des mots et les différentes catégories lexicales.

4. La distance lexicale

L'étude de la distance lexicale permet de comparer différentes œuvres par le vocabulaire qu'elles partagent et celui qui les sépare. Il s'agit de considérer le vocabulaire intégral de chacun des textes du corpus et de repérer ceux qui partagent des thèmes semblables.
L'analyse arborée ci-dessous tient compte de la distance lexicale en s'appuyant sur les occurrences (le calcul N)[2]

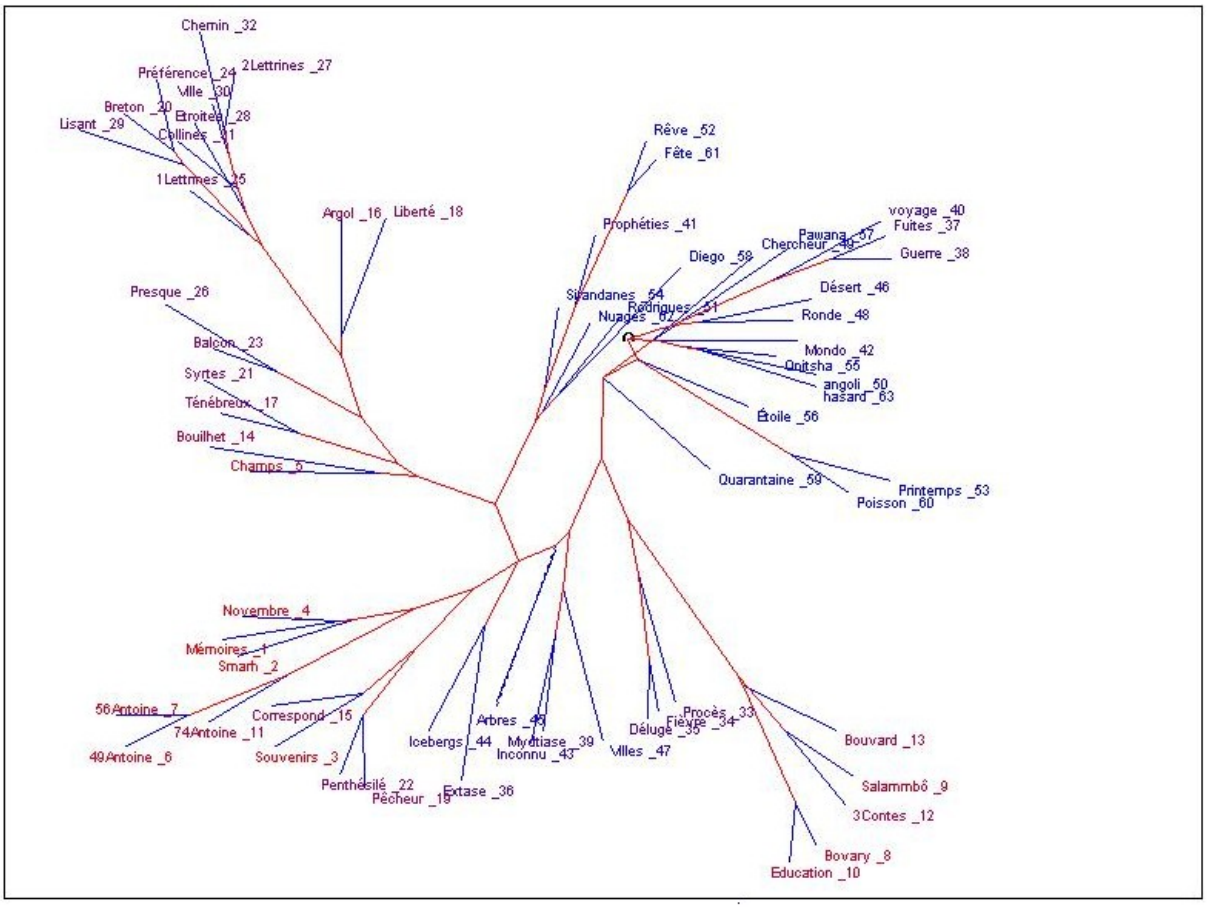


Figure n°3 : Analyse arborée de la distance lexicale s'appuyant sur les formes graphiques prenant en considération la fréquence (N)

L'arbre nous permet ici de constater premièrement la division entre les trois œuvres. L'œuvre gracquienne se trouve dans la partie supérieure gauche du graphique, l'œuvre de Flaubert dans la partie inférieure, tandis que l'œuvre leclézienne est opposée aux deux autres dans la partie supérieure droite du graphique.
En effet, il s'agit de trois écrivains très différents et ce graphique rend bien compte de ce fait. Il s'agit aussi d'époques différentes ce qui influence évidemment cette analyse.
Néanmoins, l'opposition des genres reste un facteur très important. A l'intérieur de chaque œuvre, nous pouvons bien distinguer chez Gracq les essais des romans. Quant aux pièces de théâtre *Le*

[1] Cf. M. Kastberg Sjöblom (2006) p. 108-127.
[2] E. Brunet (2006) p. 52.

Roi-pêcheur et *Penthésilée*, la différence de genre les poussent dans le camp Flaubert, tout en bas de l'arbre. Les livres de Flaubert se divisent aussi selon le genre. Le graphique met en relief les ouvrages qui partagent soit le même thème comme les trois versions de la *Tentation de Saint Antoine*, soit le même genre comme les romans *Madame Bovary*, *L'éducation sentimentale* et *Salammbô*. C'est aussi le lien du genre qui rapproche l'écriture personnelle des *Souvenirs* et de la *Correspondance*.
La structure lexicale de l'écriture leclézienne est également déterminée par le genre. Les ouvrages ethnologiques comme *La fête enchantée* et *Le rêve mexicain* ou la biographie ou les essais littéraires tardifs sont bien séparés de l'œuvre romanesque regroupée à droite du graphique. Il est toutefois intéressant de pouvoir constater que certains ouvrages de Le Clézio transgressent aussi la "frontière leclézienne" pour se trouver au milieu des ouvrages de Flaubert. Il s'agit d'un côté des essais poétiques (*L'inconnu sur la terre* et *Mydriase*) et de l'autre côté des essais littéraires écrits avant la rencontre avec le monde amérindien comme *Parmi les icebergs* et *L'extase matérielle.* Nous trouvons ici également bien regroupés les premiers livres de Le Clézio appartenant à l'école du "nouveau roman" ; *Le procès-verbal*, *La fièvre* et *Le déluge*. Ce style particulier d'écriture semble, selon cette analyse, beaucoup plus proche du style de Flaubert que celui qu'emploie Le Clézio dans ses autres ouvrages.

Conclusion

Ainsi, la numérisation et l'analyse lexicométrique de la quasi totalité des textes de trois grands écrivains français nous ont permis de mettre en exergue non seulement les changements esthétiques, mais surtout l'importance de l'opposition générique qui s'observe à tous les niveaux de l'écriture : dans la structure, dans la morphologie, dans la syntaxe aussi bien que dans le vocabulaire.
Bien que ces auteurs, et les époques soient très différents et que l'on pourrait porter à cette étude une critique sur leur comparabilité, la division générique ressort de l'analyse comme le facteur prépondérant.
Cette analyse souligne en effet la fragilité des études qui s'appuient sur la distance intertextuelle pour résoudre des problèmes d'attribution ou de datation de textes. Ces études se révèlent parfois être totalement biaisées par justement la division générique ou typologique.
Chaque genre littéraire a en fait son anatomie, sa physiologie et son fonctionnement, et cela transparaît très clairement dans les différents textes qui forment le corpus relativement hétéroclite de ces trois auteurs.

BIBLIOGRAPHIE

ADAM, J.-M., GRIZE, J.-B., & BOUACHA, M. A. 2004. *Textes et discours : catégories pour l'analyse*, Dijon, PU Dijon, Collection Langages EUD.
ADAM, J.-M. 2005. *Les textes types et prototypes : Récit, description, argumentation, explication et dialogue*, Paris, Armand Colin.
BRUNET, E. 1988. *Le vocabulaire de Victor Hugo*, Paris-Genève, Champion-Slatkine.
BRUNET, E. 2003. Flaubert traité par Hyperbase, Rouen, Revue Flaubert n° 3.
BRUNET, E. 2006. *Hyperbase, Manuel de référence*, Nice, CNRS.
CATACH, N. 1994. *La ponctuation*, Paris, PUF.
KASTBERG SJÖBLOM, M. 2006. *L'écriture de J.M.G. Le Clézio – Des mots aux thèmes*, Paris, Honoré Champion, Collection "Lettres Numériques".
KASTBERG SJÖBLOM, M. 2004. Comment l'ordinateur peut-il servir dans l'étude stylistique d'un texte littéraire et de quelle façon l'analyse de la distribution des parties du discours peut-elle contribuer à la compréhension des textes ?, in M. Ballabriga & F.-Ch. Gaudard (éds.), *Champs du Signe*, Toulouse, Editions Universitaires du Sud, pp. 119 -152.
MALRIEU, D. & RASTIER, F. 2001. Genres et variations morphosyntaxiques, in A. Martin Municio (éd.), Actas del segundo seminario de la escuela interlatina de altos estudios en linguística aplicada, Matemáticas y tratamiento de corpus, San Millán de la Cogolla*,* 19-23 septiembre de 2000, , Logroño, Fundación San Millán de la Cogolla.
RASTIER, F. 2001. *Arts et Sciences du texte*, Paris, PUF.

L'ADAPTATION COMME CONTRACTION. L'ANALYSE INFORMATISÉE DE L'ANTIGONE DE JEAN COCTEAU

Jocelyne LE BER
Collège militaire royal du Canada

SOMMAIRE

1. Le mythe d'Antigone

Antigone, à travers les siècles, a eu divers rôles à jouer. En 1580, Garnier insiste, d'après Simone Fraisse, à « rajeunir les tragédies antiques et instaure une distance immense entre le polythéisme des Grecs et le catholicisme triomphant du XVI[e] siècle finissant » (21) pour faire entendre les échos de la situation historique. Soixante ans après, en 1637, Rotrou reprend le mythe *d'Antigone* en apportant quelques nuances. Par la désobéissance d'Antigone, Raymond Trousson affirme que Routrou pensait que « l'exaltation du pouvoir absolu entraîne la négation de l'individu et la subordination des valeurs morales aux nécessités de l'organisation politique et sociale » (109). Mais c'est surtout le récit épique de Balanche, en 1814, qui fait d'elle une héroïne moderne, une sainte comparable à Jeanne d'Arc pour le dévouement et l'esprit de sacrifice[1]. C'est ainsi, que petit à petit, Antigone devient le symbole de la révolte et de la liberté anticonformiste que nous retrouvons chez Cocteau dans une mise en scène avant-gardiste et plus tard chez Anouilh, qui fait de son personnage « le symbole de la Résistance française » (110).

2. L'Antigone de Cocteau

L 'Antigone[2] de Cocteau est une contraction de la tragédie de Sophocle. Par ce procédé, Cocteau tente de redonner vie à un mythe de façon innovatrice. En contractant la tragédie de Sophocle, Cocteau a su garder le sens tragique du personnage d'Antigone qui préfère « plaire aux morts » plutôt que de « plaire aux vivants» (*TC* 308) et qui veut être à la hauteur de ses exigences éthiques, en restant digne des siens même s'ils sont dorénavant des fantômes. Elle est persuadée que ceux-ci la regardent : «les enfers et ceux qui les habitent m'ont vue agir, moi » (314). Dès l'exposition des faits, le personnage s'inscrit dans un système philosophique qui le présente et le fait se dérober à l'immanence du jugement de l'homme qui le condamnera. L'*Antigone* de Cocteau, contrairement à l'*Antigone* d'Anouilh, a déjà fait son choix au moment où le spectacle commence[3]. Elle a choisi d'épouser la mort, car dit-elle « je suis née pour partager l'amour, et non la haine » (*TC* 313), ce qui permet à Cocteau de noter que « les personnages *d'Antigone* ne s'expliquent pas. Ils agissent. Ils sont l'explication de théâtre qu'il faudra substituer au théâtre de bavardages. Le moindre mot, le moindre geste, alimente la machine » (*CJC 10* 93). La tragédie est, en ce sens, la représentation des conséquences que son acte implique puisque le choix a déjà été fait. Pierre-Aimé Touchard rappelle, à ce propos, que la conception sophocléene tendait à « concentrer

[1] En effet, l'Antigone de Ballanche est dédiée « à la duchesse d'Angoulême, fille du roi martyr, en qui, la même année, le panégyrique de Louis de Saint-Hugue saluait la nouvelle Antigone » (Trousson 109).

[2] Nous citons *Antigone* (1927) dans la récente édition de la Pléiade : *Théâtre complet*, Paris, Gallimard « Bibliothèque de la Pléiade », Edition publiée sous la direction de Michel Décaudin avec la collaboration de Pierre Caizergues, Pierre Chanel, Gérard Lieber, Francis Ramirez, Christian Rolot et Jean Touzot. 2003.

[3] Dans son prologue, Anouilh présente ses personnages de façon à ce qu'ils semblent découvrir leur rôle au fur et à mesure que la pièce se déroule. Au début de la pièce le prologue s'avance et nous dit : « Voilà. Ces personnages vont vous jouer l'histoire d'Antigone, c'est la plus maigre qui est assise là-bas, et qui ne dit rien. Elle regarde droit devant elle. Elle pense. Elle pense qu'elle va être Antigone tout à l'heure (Anouilh 35).

l'attention sur un individu et que les héros de Sophocle doivent agir de telle ou telle manière pour rester fidèle au plus profond de leur être, pour ne pas perdre la raison même de leur existence, pour maintenir leur identité au prix même de leur vie » (Touchard 41). C'est ainsi, dit-il que « la soumission au destin, chez Sophocle, n'est pas entrecoupée de révolte, elle est devenue une sagesse » (42). Toutefois, chez Cocteau l'apparente soumission d'Antigone à son destin est purement verbale et sa réticence à entrer dans la caverne comme les injonctions désespérées qu'elle lance au chœur sont les preuves de sa révolte et de sa peur de la mort. D'ailleurs, dans ses adieux, Antigone affirme « qu'on [lui] vole [sa] part de vie » (*TC* 319).

En adaptant, par la contraction, la pièce de Sophocle, Cocteau a redonné vie au mythe d'Antigone. Pour saisir l'importance de cette contraction, nous avons voulu analyser l'*Antigone* de Cocteau en la comparant à la tragédie de Sophocle à l'aide de l'informatique. Dans le présent article, nous montrerons comment Cocteau est resté « fidèle » à son prédécesseur malgré la contraction.

Cocteau disait « je déblaye, je concentre le texte de Sophocle ». Il explique cette contraction dans la préface de la pièce :

> C'est tentant de photographier la Grèce en aéroplane. On lui découvre un aspect tout neuf. Ainsi, j'ai voulu traduire *Antigone.* À vol d'oiseau, de grandes beautés disparaissent, d'autres surgissent; il se forme des rapprochements, des blocs, des angles, des reliefs inattendus. (*TC* 305)

Dans son article « Jean Cocteau Early Greek Adaptations», Carol A.Cujec, note qu'en réduisant le texte, Cocteau en aurait écarté son aspect lyrique, « ce qui lui permet d'expérimenter les pouvoirs expressifs des nombreuses ressources théâtrales afin de créer son style unique de la poésie de théâtre [1] (47). Simone Fraisse, de son côté, affirme que Cocteau « avait conscience d'avoir débarrassé la tragédie antique de ses poussiéreuses bandelettes » (116). La nouveauté, selon elle, ne serait pas vraiment dans le texte, puisque Cocteau aurait simplement réduit la tragédie de moitié. L'informatique permet de préciser les affirmations de Simone Fraisse en même temps que l'aspect thématique de la contraction. Afin de procéder à une comparaison de deux écrivains, nous avons créé un corpus numérisé, corpus que nous avons ensuite traité avec les logiciels *grep*[2], *freq*[3] et *awk*[4] . Nous avons ainsi tenté de distinguer les points communs et la divergence des deux œuvres. Dans un premier temps, nous avons donc repris les textes de Sophocle et de Cocteau et nous les avons étiquetés de façon à pouvoir séparer les répliques des personnages et de compter les mots qui leur étaient attribués dans la pièce. Dans un deuxième temps, nous avons également étudié les différences thématiques et stylistiques entre les deux pièces.

3. L'analyse quantitative

L'informatique nous permet d'être plus précis concernant la réduction du texte de Sophocle. C'est à l'aide des logiciels *grep*, *freq* et *awk* que nous avons pu identifier les parties du texte qui avaient été contractées. Les tableaux qui suivent nous permettront d'expliquer plus en détail les contractions des répliques attribuées aux personnages principaux que sont le Chœur, Créon et Antigone. Tous les calculs sont chiffrés en pourcentage. Le premier tableau représente le pourcentage des répliques attribuées à Antigone et à Créon dans les deux textes. Nous remarquerons que Cocteau attribue presque autant de répliques à Antigone que l'avait fait Sophocle. Si l'on fait la somme des répliques d'Antigone et de Créon, Antigone a 38% de ces répliques chez Sophocle alors qu'elle en a 35% chez Cocteau. En termes absolus, on trouve 46 répliques pour l'Antigone de Sophocle comparées à 45 répliques chez Cocteau.

---

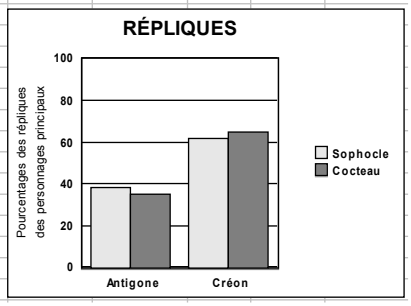
[1] Traduit de : « Allowing him to experiment with the expressive powers of the many other theatrical resources to create his unique style of "Poésie de Théâtre" » (Cujec 47).

[2] Le Grep est un procédé de Recherche/sélection/remplacement où, au lieu de chercher une chaîne donnée, on cherche une chaîne dans le document qui correspond à un motif donné, selon des règles précises, en partant du début du document. On obtient, le cas échéant, une sélection continue. Le remplacement s'effectue alors encore d'après un autre motif, toujours selon des règles bien précises.

[3] Logiciel qui permet de calculer les occurrences.

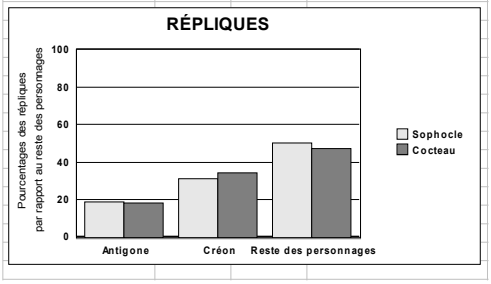
[4] Logiciel qui permet le traitement des flux de données.

	Antigone	Créon
Sophocle	38	62
Cocteau	35	65



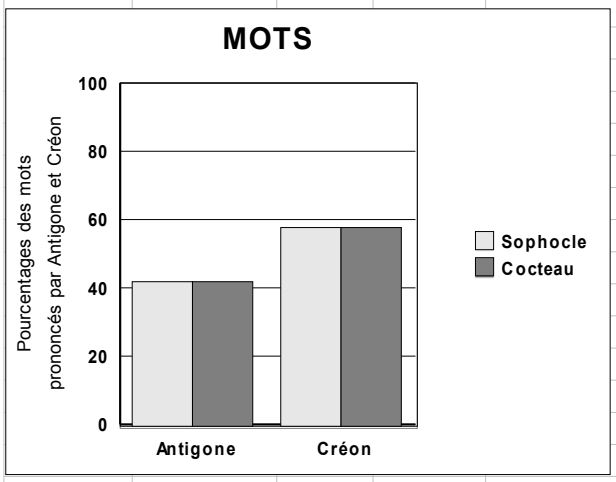
Pour Créon, nous sommes arrivée à des conclusions semblables quant au peu de différence entre la proportion des répliques attribuées aux deux personnages principaux, bien que cette fois c'est Créon qui ait légèrement plus de répliques chez Cocteau : 62% chez Sophocle et 65% chez Cocteau (soit en termes absolus 75 répliques pour le Créon de Sophocle et 84 répliques chez Cocteau). Nous nous sommes également demandée si la distribution des autres rôles avait été respectée chez Cocteau. Le tableau ci-dessous montre qu'Antigone a 19% de toutes les répliques dans la pièce de Sophocle et 18% dans celle de Cocteau. Les répliques de Créon correspondent à 31% dans l'ensemble chez Sophocle et 34 % chez Cocteau. Les autres personnages de la tragédie se répartissent 50% de la pièce chez Sophocle et 47% chez Cocteau. Nous constatons donc peu de différences entre Sophocle et Cocteau en ce qui concerne le nombre de répliques pour l'ensemble des personnages.

	Antigone	Créon	Reste des personnages
Sophocle	19	31	50
Cocteau	18	34	47



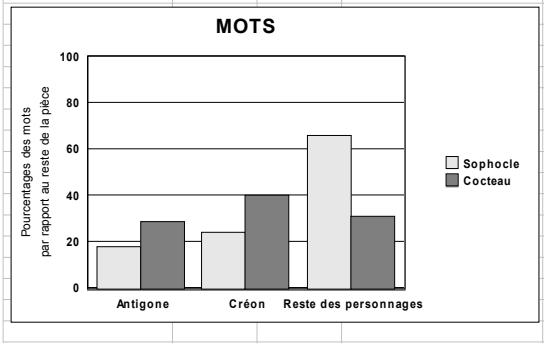
Considérer seulement la répartition des répliques ne nous semble pas suffisant. Il faut également analyser le nombre de mots attribués à Antigone et à Créon. Nous avons donc séparé et compté, à l'aide du logiciel *grep*, les mots des répliques d'Antigone et de Créon. Nous avons obtenu le même pourcentage chez Sophocle et chez Cocteau, comme le montre le tableau suivant. La fréquence absolue des mots chez Antigone de Sophocle est de 1683 (42%) et de 2312 (58%) chez Créon. Chez Cocteau la fréquence absolue d'Antigone est 1478 (42%)et de 2069 (58%) chez Créon.

	Antigone	Créon
Sophocle	42	58
Cocteau	42	58



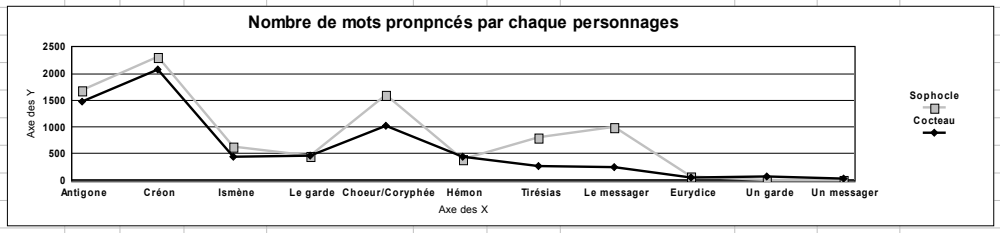
Par contre, lorsque nous avons calculé le nombre de mots de chaque personnage par rapport au reste du texte, le résultat était nettement différent, comme le montre le tableau suivant. En effet, les répliques d'Antigone représentent (1683) 18% du texte de Sophocle par rapport à (1478) 29 % chez Cocteau. Créon récite (2312) 24 % du nombre de mots chez Sophocle et (2069) 40% des mots chez Cocteau. Le reste des personnages se partagent 66% du texte chez Sophocle et 31% chez Cocteau.

	Antigone	Créon	Reste des personnages
Sophocle	18	24	66
Cocteau	29	40	31



La différence que nous constatons dans ce tableau est due à la présence du chœur et du Coryphée chez Sophocle. Dans le tableau ci-dessous, nous remarquons la relative fidélité de Cocteau par rapport au texte de Sophocle en ce qui concerne la place qu'occupent les personnages principaux. Le graphique montre bien où s'est effectuée la contraction du texte. Nous pouvons donc conclure à la vue de ce tableau que c'est bien au niveau du chœur, de Tirésias et du Messager que s'est effectuée la contraction de Cocteau.

	Antigone	Créon	Ismène	Le garde	Choeur/Coryphée	Hémon	Tirésias	Le messager	Eurydice	Un garde	Un messager
Sophocle	1683	2312	645	470	1616	396	816	997	83	0	0
Cocteau	1478	2069	440	467	1016	435	265	238	52	69	32



4. L'analyse qualitative

Cocteau n'a pas seulement réduit de moitié le mythe d'Antigone, il y a aussi selon Simone Fraisse le style que Cocteau aurait rendu « plus nerveux, plus familier pour accuser les heurts entre les personnages » (Fraisse 117). Toutefois, elle ajoute ensuite qu'à « travers cette apparente fidélité, Cocteau laissait percer ses propres choix » (Fraisse 117). Cocteau ferait d'Antigone un être amoral, comme le suggère l'épigraphe du *Voyage de Sparte* de Barrès qu'il a mise en exergue de sa pièce :

> Je pleure Antigone et la laisse périr.
> C'est que je ne suis pas poète.
> Que les poètes recueillent Antigone. — Voilà
> Le rôle bienfaisant de ces êtres.
> (Maurice Barrès)

Ainsi, Cocteau aurait « insufflé sa haine de l'ordre et son goût de l'anarchie. Il met en valeur ce dernier mot alors que jusque-là les traducteurs préféraient désobéissance » (Fraisse 117) Il est vrai que l'on pourrait supposer que Cocteau préfère l'anarchie à l'ordre. À ce sujet, le poète s'explique, en octobre 1925, dans une lettre adressée à Jacques Maritain :

> Je me confesse. J'ai laissé mon civisme en friche; ma justice se forme donc sans parti pris. Or l'instinct me pousse toujours contre la loi. C'est la raison secrète pour laquelle j'ai traduit *Antigone*. Je détesterais que mon amour de l'ordre bénéficiât du sens que l'on prête paresseusement à ce mot (Cocteau 284).

Pourtant si Simone Fraise relève la préférence des traducteurs pour le mot « désobéissance », il est important, pour nous, de considérer les différentes possibilités, car selon Georges Steiner :

> le traducteur est le facteur de la pensée et des sentiments humains. À tous les carrefours temporels et spatiaux, les flux énergétiques de la civilisation sont véhiculés par la traduction, par le processus d'échange mimétique, adaptateur et métamorphique du discours et des codes. (Steiner 220).

Ainsi, dans les traductions que nous avons consultées, nous avons trouvé différentes versions. Le sens premier reste le même, mais les connotations changent. Dans une traduction nous retrouvons le mot "désobéissance" (refus d'obéir), et dans l'autre "rébellion" (action de révolte), mais nous avons également trouvé le mot "anarchie[1]".

> **Sophocle**
> Qu'elle implore Zeus protecteur des liens du sang: si je souffre la *désobéissance* dans mes proches, que dois-je attendre des étrangers ? [...] L'*anarchie* est le plus grand des maux : elle ruine les cités, bouleverse les familles, jette les armées dans le désordre et la fuite; ceux, au contraire, qui restent fermes à leur poste, l'obéissance fait leur salut (Sophocle,1998,62)[2].
> **Sophocle**
> Là dessus qu'elle chante patenôtres à Zeus garant des liens du sang ! Si je laisse prospérer la *rébellion* dans ma famille, j'aurai du beau avec ceux qui ne sont pas ! [...] Mais l'arrogant qui

[1] Dans les citations qui suivent, c'est nous qui soulignons les mots.
[2] Sophocle. *Antigone*. Traduction de Aziza Claude. Pocket classique. France : Paris, 1998.

> viole les lois ou prétend donner des ordres à qui gouverne, en aucun cas il ne recevra mon approbation. [...] Il n'est pire fléau que le *refus de l'autorité* (Sophocle 1999, 434)[1].

Cocteau utilise également une fois le terme "anarchie" et deux fois le terme "anarchiste(s)": « Il n'y a pas de plus grande plaie que l'*anarchie*. Elle ruine les villes, brouille les familles, gangrène les militaires. Et si l'*anarchiste* est une femme, c'est le comble. Il vaudrait mieux céder à un homme. On ne dira pas que je me suis laissé mener par les femmes. (*TC* 315), ainsi que : « c'est donc bien agir que de louer les *anarchistes*» (*TC* 317). Simone Fraisse n'est pas la seule critique à relever cet "anachronisme". Carol A. Cujec le relève également dans son analyse du chœur. Il soutient que la tragédie d'*Antigone* serait pour Cocteau une occasion d'enlever d'un texte vieux de plusieurs siècles tout ce qui le ralentissait, donc tout ce qui empêchait le spectateur de communiquer avec le texte. Il aurait réduit le dialogue à son simple usage :

> En simplifiant et en éliminant la rhétorique de Sophocle, [Cocteau] a créé un texte auquel le public moderne réagit plus aisément. Il se permet même à l'occasion un anachronisme quand Créon appelle Antigone une anarchiste. (Cujec 47)[2]

Or, le mot « anarchie » dans le texte de Cocteau ne constitue pas un anachronisme, tels que le suggéraient Fraisse et Cujec, puisque le mot « anarchie » existe depuis l'Antiquité et vient du grec *anarchia*. Il n'en est pas moins un mot important dans la pièce de Cocteau, qu'il faut considérer dans le réseau des termes associés à la révolte. En fait, l'analyse informatique, et plus particulièrement le logiciel *diff*, a permis de déceler plusieurs mots qui figurent dans le texte de Cocteau sans figurer dans le texte de Sophocle. La liste de mots du texte de Cocteau comprend plusieurs noms, adjectifs et verbes utilisés dans un contexte de révolte : *anarchiste, antipatriotique, barricade, conciliabule, corrompu, despotisme, envahisseur, gangrène, militaire, résistance, révolte, terrorise*, *torture*. La France à l'époque de l'adaptation d*'Antigone* est dans un bouleversement économique et culturel, ce qui expliquerait cette tentation d'employer ce genre de vocabulaire. Mais cette « anarchie de salon » ne met pas l'ordre en danger. Toutefois, le fait que Cocteau utilise trois fois le mot « anarchie » dans son texte nous permet de suggérer que Cocteau invitait les spectateurs de l'époque à réfléchir sur les conséquences d'un pouvoir absolu.
D'après notre analyse informatique, il n'y a qu'un véritable anachronisme dans l'adaptation du mythe d'Antigone de Cocteau. En effet, dans une réplique d'Antigone, le mot « voiture » est mentionné une fois lorsqu'elle s'adresse aux Thébains : « Moquez-vous de moi; c'est bien le moment ; je vous le conseille. Ils n'attendent même pas que je disparaisse ! Ah ! Thèbes ! Ah ! Ma ville aux belles voitures ! » (*TC* 318). Toutefois, si nous considérons le poids de l'anachronisme dans le texte de Cocteau, il est certain que le mot « voiture » n'a pas la même importance qu' «anarchiste». Par contre, le mot «voiture» est un anachronisme qui permet d'actualiser la pièce, tout comme le fera, 20 ans plus tard, Anouilh, qui fera par ailleurs d'Antigone l'emblème de la révolte et de la résistance au moment de l'Occupation.

4.1. La contraction du chœur

Comme nous l'avons montré, la contraction du texte de Sophocle s'opère surtout au niveau du chœur. À l'aide de l'informatique, nous avons pu séparer toutes les répliques du chœur et du Coryphée du texte de Sophocle et du chœur de Cocteau. Cette opération s'est révélée très intéressante, car, lorsque nous nous sommes penchée sur la comparaison des répliques, il s'est avéré que Cocteau avait plus particulièrement utilisé le rôle du Coryphée dans ses répliques du chœur. En effet, il efface des dialogues tout le côté noble et sublime du langage poétique propre à Sophocle. Il rend le vocabulaire plus familier et resserre le discours. Comparons, à titre d'exemple, ces deux répliques :

---

[1] Sophocle. *Les Tragiques grecs, Théâtre complet*. Traduction et note de Débidour Victor-Henri. Éditée avec une introduction générale et un dossier sur la tragédie par Paul Demont et Anne Lebeau. Le livre de Poche. France : Paris, 1999.
[2] Traduit de : « He even allowed for the occasional anachronism, as when Créon calls Antigone an anarchist.».

Sophocle :
Sur le bon et sur le mauvais serviteur du pays Créon, fils de Ménécée la sentence est rendue, c'est bien : il t'appartient de porter des décrets à ta guise aussi bien sur les morts que sur nous autres les vivants (Sophocle 1999, 422).
Cocteau :
Bravo, Créon. Tu es libre, tu disposes des morts et de nous (*TC* 309).

Nous pouvons dans ces deux passages voir la réduction du texte. Cocteau utilise un style plus direct que Sophocle et peut-être ironique à cause du « bravo »[1]. La contraction se base également sur l'élimination de la périphrase qui sillonne le texte de Sophocle. Ainsi, dans ces citations du chœur s'adressant à Antigone :

Sophocle :
Glorieuse, admirée, tu t'en vas vers ce monde secret où sont les morts. Ni une maladie ne t'a flétrie, ni une épée ne t'a meurtrie : prenant ta loi en toi-même vivante, Ô destin inouï, tu vas descendre chez Hadès. (Sophocle 1999, 438).
Cocteau :
Tu mourras donc sans être malade, sans blessure. Libre, vierge, vivante, célèbre, seule entre les mortels, tu entreras chez Pluton *(TC* 318)

Dans cet exemple, Cocteau effectue un résumé de l'action en employant un adjectif ou un champ sémantique à la place des périphrases de Sophocle: sans être malade ("ni une maladie ne t'a flétrie"), sans blessure ("ni une épée ne t'a meurtrie"), libre, vierge, vivante ("prenant ta loi en toi-même vivante"), célèbre ("glorieuse, admirée"). Par ce refus de l'amplification, Cocteau permet au chœur de donner une simple énumération des qualités d'Antigone. Cet autre exemple, où Cocteau "contracte" une tirade du chœur, mérite d'être cité. Nous avons choisi de faire alterner les répliques de Sophocle et celles de Cocteau pour bien illustrer la réduction systématique des périphrases métaphoriques :

Sophocle
Mille prodiges par le monde... Mais l'Homme est le plus haut prodige : il passe la mer écumeuse, le vent du sud, en ses bourrasques, le porte : il passe au creux des lames qui se gonflent et le cernent de leur abois. Et la Terre, divine et toute souveraine, impérissable, intarissable d'année en année, il l'éventre au va-et-vient de ses charrues ou il attelle les bêtes dont il a peuplé ses écuries (Sophocle 1999, 425).
Cocteau
L'homme est inouï. L'homme navigue, l'homme laboure (*TC* 311).
Sophocle
Les oiseaux à l'âme légère dans ses rets il les enveloppe ; les ordres des bêtes sauvages et la faune océane en mer il les capture au fond des mailles du filet qu'il sait tresser, dans son astuce, lui l'Homme ! Et ses engins maîtrisent l'animal qui gîte aux champs, qui court les monts ; sous le joug qui serre leur nuque le cheval offrira son col empanaché le taureau montagnard son inlassable effort (Sophocle 1999, 426).
Cocteau
L'homme chasse, l'homme pêche. Il dompte les chevaux (*TC* 311).

Cette concision permet également au chœur d'être plus radical. Malgré ces changements, contrairement aux hypothèses formées par Carol. A. Cujec, pour qui les contractions rendent « le chœur et Créon plus haïssables que dans Sophocle » (Cujec 49), le rôle du chœur ne change pas; il continue plutôt, comme le dit Hegel au sujet de la tragédie antique, dans son ouvrage intitulé *Esthétique*, à exprimer « des idées et des sentiments généraux » (Hegel 342). Par ailleurs, Cocteau, lui-même, affirme que le chœur peut changer d'opinion et être ainsi sensible aussi bien aux problèmes de Créon qu'à ceux des autres personnages de la tragédie.

4.2. La contraction des répliques de Créon

Le Créon de Cocteau, comme le Créon de Sophocle, est la raison d'État. Politiquement, il est parfaitement plausible que sa décision soit d'intérêt public. Il ne comprend pas Antigone et, au fur

[1] Ce qui pourrait expliquer les rires dans la salle lors des représentations.

et à mesure, ce n'est plus par devoir, ni même par son autorité, mais par colère, par dépit, par inintelligence butée et rageuse, qu'il raidit sa condamnation. Tout ce qui devrait le rappeler à la générosité qu'il se doit : dignité d'Antigone, noblesse du cœur d'Ismène, protestation d'Hémon, lamentations du chœur et de la victime avant le supplice, tout cela durcit son parti pris.
Ce qui distingue la tragédie de Sophocle et celle de Cocteau est le niveau de langage. Sophocle permet à Créon de s'exprimer, tout comme nous l'avons vu avec le chœur, à l'aide de périphrases, ce qui donne à Créon une posture princière. Par contre, Cocteau, afin d'actualiser sa pièce et de la rendre plus accessible, écrit pour Créon des répliques brèves et sèches.
Ainsi, lorsqu'il s'adresse aux Thébains et qu'il décrète sa loi, il finit sa réplique par « j'ai dit» *(TC* 309*).* Cette façon de terminer son discours permet de supposer d'emblée, dès son arrivée sur scène, que Créon tiendra un rôle d'autorité. Il confirme sa puissance et le spectateur comprend qu'il n'acceptera aucune revendication. Il y a donc ici le stéréotype du dictateur. Nous constatons également qu'il y a une « fidélité » au rôle de Créon . Cocteau, malgré sa "contraction", reprend les connotations reliées à l'autorité. Comparons ces deux exemples :

> **Sophocle**
> Mais il y a des gens, dans la ville, qui d'emblée ont *renâclé* contre mes ordres ; *murmures*, hochements de tête à la dérobée, refus de plier la nuque sous le *joug* légitime et de prendre en gré *mon autorité.* Ce sont eux, je le sais bel et bien, qui ont séduit le piquet de garde et l'ont payé pour faire ce travail-là... Chez les hommes, nulle puissance établie n'a commis plus de méfaits que *l'argent*. C'est lui qui fait la ruine des États, lui qui chasse les gens de leur foyer, lui dont les leçons égarent les consciences vertueuses en les incitant à des infamies. C'est lui qui a appris aux humains les scélératesses, et le secret de tous les actes impies ! (Sophocle, 1999, 424)
> **Cocteau**
> [...] Je savais déjà que des *traîtres murmurent* contre mon *joug* dans cette ville, qu'on se soulève en cachette. Ils payent les coupables. Les mortels ont inventé *l'argent. L'argent, l'argent* ignoble ! *L'argent* ruine les villes, fausse les cœurs droits, démoralise tout. *(TC* 310-11).

Cocteau reprend les mots clefs de l'autorité. Il garde le mot "joug", qui d'après le *Petit Robert*, est « une contrainte matérielle ou morale qui pèse lourdement sur la personne qui la subit, entrave ou aliène sa liberté », il garde également l'idée que l'argent est un mal pour la société. Par contre, il remplace le verbe d'action « renâcler », qui signifie : «témoigner de la répugnance », par « le murmure des traîtres », qui n'a plus une valeur de répugnance pour le pouvoir, mais une valeur qui suggère un mouvement de révolte chez le peuple.
Contrairement à Sophocle, Cocteau ne laisse pas les spectateurs s'émouvoir par le rôle de Créon, qui inspire plutôt l'antipathie par la sévérité de ses paroles.

> **Sophocle :**
> Ô toi qui tiens les yeux baissés vers la terre, avoues-tu, ou nies-tu avoir fait ce dont il t'accuse ? (Sophocle, 1999, 428)
> **Cocteau :**
> Et toi. Toi, avec tes yeux modestes, tu nies ? Tu avoues ?
> *(TC* 312*)*

Le fait que Cocteau n'emploie pas l'interjection «Ô», traduisant un vif sentiment, qu'il inverse dans la phrase les verbes «nier» et «avouer», et qu'il interpelle Antigone par le «et toi» nous indique encore la rudesse du personnage et un caractère quelque peu roturier. Le «et» précédant le pronom personnel «toi» rend le dialogue plus familier et permet à Cocteau de transformer l'invocation en une interpellation en dirigeant le discours de Créon vers Antigone. Cocteau exploite également les métaphores et l'exclamation pour renforcer la colère de Créon.

> **Sophocle**
> Mais les esprits inflexibles s'abattent aisément: *le fer le plus dur s'amollit par la flamme et se brise; un léger frein réprime la fougue des plus fiers coursiers*. L'orgueil sied mal à l'esclavage. Elle savait qu'elle m'outrageait en violant mes ordres; elle ajoute à son crime celui d'en tirer vanité et de sourire avec dédain. (Sophocle, 1999,429).

Cocteau
Mais sache que ces âmes si dures sont les moins solides. *C'est le fer le plus dur qui éclate. Un petit mors calme un cheval qui fait des siennes***.** Voilà beaucoup d'orgueil pour une esclave...Une esclave du devoir. Elle m'outrage exprès. Elle me nargue et s'en vante. (*TC* 312).

Nous avons dans cette réplique de Créon, par la métaphore d'un cheval fougueux difficile à dompter, la métaphore du caractère bouillant d'Antigone. Créon ne peut supporter cette impétuosité, il aimerait pouvoir la dresser : « un petit mors calme un cheval qui fait des siennes ». Il veut assujettir Antigone qu'il la compare à « une esclave du devoir » qui doit obéir à son maître. Le Créon de Sophocle n'était pas aussi radical dans ses répliques. Le sens est le même, mais le langage plus noble ne nous permettait pas d'envisager le caractère colérique de Créon. Le Créon de Sophocle est plus général dans ses propos, donc paraît plus diplomate. Il nous semble important de signaler par d'autres exemples cette radicalisation que Cocteau opère dans les répliques de Créon, car Cocteau, en utilisant un style plus direct, fait ressortir la sévérité de Créon. Nous nous proposons, à titre d'exemple, une partie du dialogue entre Créon et son fils Hémon :

Sophocle
Créon : Comment, misérable ? en mettant ton père en accusation?
Cocteau
Créon : Canaille! Tu insultes ton père.
Sophocle
Hémon : C'est que je te vois t'égarer dans les iniquités.(Sophocle, 1999, 436)
Cocteau
Hémon : C'est que je vois mon père injuste. *(TC* 316*)*

Nous constatons dans ces exemples l'emploi du tutoiement. D'après Cujec :

> Cocteau se distance du texte; ce qui a trait au discours familier en réanimant l'usage d'anciennes pratiques : l'emploi exclusif du pronom personnel "tu" tout au long des répliques des personnages[1][...] (Cujec 47)

Cujec a montré qu'avec l'emploi du "tu" il y a eu un écart par rapport à la langue courante où, dans un contexte d'autorité, on se serait attendu à un vouvoiement. Selon nous, l'usage du "tu" chez Cocteau, ajouté à la contraction du texte, semble soutenir également le style direct de Créon.

4.3. La contraction des répliques d'Antigone

Antigone, chez Sophocle, est une jeune fille élevée dans le bonheur d'un foyer princier. La révélation de sa filiation monstrueuse, le suicide de sa mère, la mutilation de son père ont ravagé sa conscience. À présent la mort de ses deux frères a emporté « ce qui pouvait lui rester d'esprit d'abandon, de docilité à la Providence. Ce n'est plus qu'une énergie à vif, qui ne sait sur quoi se porter, mais qui, dès qu'elle l'aura trouvé, s'y portera avec une sorte de lucidité aveugle. En se sens, elle prend la suite de cette condamnation héréditaire à laquelle, elle se sait, bien plus, à laquelle elle veut se vouer : elle se construit une liberté de révolte et de sacrifice, mais c'est à l'intérieur du cachot qui emprisonne sa race. Elle aussi, « la dernière et la plus misérable », elle paiera pour Œdipe » (Lebeau 411). Cocteau garde cette image de jeune fille révoltée, libre et prête à se sacrifier pour une cause. Il donne également à Antigone l'effigie de la sainte. En effet, le personnage d'Antigone nous paraît pur par son action envers son frère et transparente par son honnêteté. Le choix d'une adaptation d'Antigone ne nous paraît pas être un choix arbitraire, car elle semble incarner l'esprit indépendant et de rébellion contre un ordre social établi dans la loi de la poésie. La déclaration d'Antigone à Ismème « je sais que je plais où je dois plaire » *(TC* 308*),* peut-être interprétée comme un cri de bataille contre ses ennemis. Antigone est punie par la société, ce qui fait d'elle la première martyre dans l'œuvre de Cocteau[2]. Pourtant, bien que

[1] Traduit de : « Cocteau distances the text with respect to normal speech by reviving ancient practices: the exclusive use of the informal address *tu* among all the characters[...]» (Cujec 47).
[2] On pourrait aussi rappeler ici que Cocteau se considérait constamment jugé aussi bien pour son art que pour son homosexualité. Cocteau croyait que, comme Antigone, il était condamné de son vivant à souffrir et

Cocteau voie Antigone comme une sainte, il n'exploite pas l'opportunité d'insister sur la figure de la « vierge ». Il maintient l'esprit de Sophocle pour son héroïne et garde le stéréotype de la jeune fille pure qui se sacrifie, et qui sera l'élément déclencheur pour ses œuvres futures, voire un leitmotiv.

> **Sophocle**
> Voyez-moi, citoyens de la terre ancestrale faire mes derniers pas, pour la dernière fois regarder l'éclat du soleil... Plus jamais... Toute vive, il m'entraîne, le dieu d'Enfer, en son universel Sommeil, aux bords du fleuve de Chagrin. Je n'aurai point connu les rites d'épousailles, nul cantique pour moi ne fut encore chanté sur le seuil nuptial et mes noces seront au fleuve de Chagrin ! (Sophocle 1999, 438)
> **Cocteau**
> Citoyens de ma patrie, regardez-moi. Je commence mon dernier voyage et je regarde une dernière fois la lumière du soleil. Le dieu infernal va me prendre vivante, sans que je connaisse le mariage, sans que les chants du mariage répètent mon nom; c'est la mort qui m'épouse (*TC* 317-18).

Chez Cocteau tout comme chez Sophocle, Antigone passe de "mortelle " à "immortelle" par ses "épousailles" avec la mort et s'approche d'une spiritualité authentique de l'amour et du sacrifice. Même si Antigone ne cesse de se référer, chez Sophocle, aux puissances du monde souterrain, c'est d'en haut qu'elle est appelée, car

> elle était divine et de race divine et nous sommes mortels, de race périssable... Ah ! Quand tu seras disparue, du moins auras-tu l'ample gloire d'avoir eu en partage un sort qui fut divin – vivante, et jusque dans la mort! (Sophocle 1999, 438)

Si Cocteau a donné à Antigone la figure de la vierge, il n'a toutefois pas oublié de faire valoir l'image de la liberté. Il donne à son héroïne le mandat de dévoiler le danger d'un gouvernement totalitaire. Ainsi, il se permet d'approfondir l'idée de Sophocle en permettant à Antigone d'utiliser le mot despotisme : « Alors, pourquoi traîner ? Tu me déplais et je te déplais. Toute cette foule m'applaudirait sans la crainte qui paralyse la langue. À mille autres privilèges, le despotisme ajoute celui de dire et d'entendre ce qu'il veut *(TC* 313). Nous avons vu que Créon avait le stéréotype d'un homme autoritaire qui ne pouvait supporter d'être contredit. Sophocle exploite l'image de la tyrannie, qui est reliée à l'Histoire antique, en donnant à la royauté le pouvoir d'agir à sa guise. Par contre, Cocteau, lui, accuse le pouvoir d'autorité suprême. Par cette réplique, Antigone devient l'emblème de la liberté . Elle lutte contre l'injustice et devient le symbole de la révolte. De «sainte» à révolté, elle devient martyre en se sacrifiant. Mais elle ne peut se sacrifier sans se donner une raison :

> Je vais revoir mon père, ma mère, Étéocle. Quand vous êtes morts je vous ai lavés, je vous ai fermé les yeux. Je t'ai aussi fermé les yeux Polynice et j'ai eu raison. Car jamais je n'aurais fait cet effort mortel pour des enfants ou un époux. Un époux, un autre peut le remplacer. Un fils, on peut en concevoir un autre. Mais comme nos parents sont morts, je ne pouvais espérer des frères nouveaux. C'est en vertu de ce principe que j'ai agi, qu'on me frappe, que Créon me prive du mariage et de la maternité. (*TC* 319).

Dans ces exemples, l'héroïne de l'amour fraternel et de l'abnégation s'est changée en une froide raisonneuse, qui donne un mobile à son acte, un calcul, en vertu duquel un mari, des enfants sont "remplaçables", tandis qu'un frère ne l'est pas. Antigone, par ces répliques, nous montre à quel point elle représente l'amour absolu.

Conclusion

Dans *Le Rappel à l'ordre*, Cocteau a dit qu'« un artiste original ne peut pas copier. Il n'a donc qu'à copier pour être original. » (37). Ce paradoxe rend bien compte de son *Antigone,* tel que nous l'a démontré l'analyse informatique et littéraire. En effet, l'originalité de la contraction de Cocteau, par des effets de style direct et familier, apparaît comme une réduction de la pièce de Sophocle tout en reprenant sa structure. Toutefois, Cocteau démontre aussi son originalité par sa mise en scène, qui renforce la pureté de la contraction. En effet, Cocteau a innové sur le plan des décors,

---

qu'il passerait le seuil du respect et de l'admiration seulement après sa mort.

des costumes, de la musique et du jeu des acteurs. De cette façon, il a pu donner à la tragédie d'*Antigone* une nouvelle texture pour le spectateur moderne. Malgré qu'il puisse avoir été accusé d'imitation, ces œuvres démontrent que la copie ou dirait-on maintenant l'intertextualité, n'empêche pas l'innovation. Les innovations de Cocteau encourageront d'autres dramaturges comme Anouilh à suivre et à considérer *Antigone* moins comme une tragédie religieuse, que comme une tragédie morale et politique. C'est ce que nous démontre l'*Antigone* de Cocteau, première pièce mythologique d'une longue série au XX^{e} siècle.

BIBLIOGRAPHIE

Antigone (1927) dans la récente édition de la Pléiade : *Théâtre complet*, Paris, Gallimard, Bibliothèque de la Pléiade, Édition publiée sous la direction de Michel Décaudin avec la collaboration de Pierre Caizergues, Pierre Chanel, Gérard Lieber, Francis Ramirez, Christian Rolot et Jean Touzot. 2003.
ANOUILH, J. 1946. *Antigone*. Paris, La table ronde.
BALLANCHE, P. S. 1814. *Antigone*. Paris, Didot.
COCTEAU, J. 1926. *Lettres à Jacques Maritain*, Paris, Stock.
CUJEC Carol, A. 1996. Jean Cocteau's Early Greek Adaptations, *Classical and Modern Literature*, Volume 17, Numéro 1, Automne 1996, pp. 45-56.
FRAISSE, S. 1974. *Le Mythe d'Antigone*, Paris, Armand Colin.
GARNIER, R. 1952. *La Troade*, *Antigone*, Texte établi et présenté par Raymond Lebègue Paris, Les Belles Lettres.
HEGEL, F.W. 1832. *Esthétique*, Traduction de S. Jankélevitch, Paris, Aubier-Montaigne, 1965.
SOPHOCLE. *Antigone*. Traduction de Aziza Claude, Paris, Pocket classique, 1998.
SOPHOCLE. *Les Tragiques Grecs, Théâtre complet*. Traduction et note de Débidour Victor-Henri. Éditée avec une introduction générale et un dossier sur la tragédie par Paul Demont et Anne Lebeau, Paris, Le livre de Poche, 1999. (Version utilisée pour l'analyse informatique).
STEINER, G. 1986. *Les* Antigones, Traduit de l'anglais par Philippe Blanchard, Paris, Gallimard.
TOUCHARD, P.-A. 1968. *Le Théâtre et l'angoisse des hommes*, Paris, Seuil.
TROUSSON, R. 1981. *Thèmes et mythes*, Bruxelles, Éditions de l'Université de Bruxelles.

QUEL BALISAGE POUR LES CORPUS ÉPISTOLAIRES NUMÉRIQUES ?

Françoise LERICHE
Université Grenoble 3 / ITEM (CNRS)

SOMMAIRE

1. Enjeux, propositions

Malgré les Cassandre du monde éditorial français, qui voient dans la concurrence de l'édition électronique la fin de l'édition (sélectivité, éthique, etc.), tirant argument de quelques entreprises vouées – en effet – à la diffusion hâtive et purement mercantile de textes mal édités (« édition sans éditeurs », lit-on çà et là), on peut de manière générale, en prenant un peu de recul, estimer qu'au contraire, les possibilités numériques sont en train d'introduire un nouvel essor et une mutation positive dans le domaine de l'édition et la philologie – comme l'imprimerie, à la Renaissance, avait permis l'apparition de nouveaux corpus et d'une nouvelle philologie (Rastier, 2001, p. 53 sqq.).

Parce que l'édition numérique permet de nouveaux modes d'accès aux textes (par rapport à la lecture linéaire du rouleau antique et même par rapport aux accès tabulaires du codex ou du livre imprimé – sommaires, index, etc.), la « philologie numérique » (Rastier, 1991, pp. 53-65), par diverses formes de balisages, facilite de manière inédite la lecture des textes et, mutation essentielle, révolutionne leur appréhension par la constitution de ces textes en « corpus » raisonnés. Mais le plus révolutionnaire (à mes yeux, du moins) réside dans le fait que l'ère numérique rend possible (est déjà en train de rendre possible) l'édition de textes qui résistaient à l'édition imprimée : fonds de manuscrits, vastes correspondances, œuvres multimédia, œuvres fictionnelles anciennes couvrant des milliers, voire des dizaines de milliers de pages… Ensembles textuels dont *seule* une édition électronique peut établir le texte, l'annotation, les modes d'accès et de recherche adéquats.

1.1. L'édition électronique et l'émergence de nouveaux corpus textuels et/ou documentaires

Les manuscrits[1] constituent un exemple typique de corpus laissés de côté par l'édition papier : fragmentaires par nature, raturés, offrant un ordre de lecture incertain voire incompréhensible, ils se refusent à la lecture linéaire qu'implique – malgré tout – le livre imprimé. Idéologiquement ils ont aussi, pendant des décennies, été négligés par la philologie classique, relégués à un vague statut de «documents », de traces documentaires, et non considérés comme des textes à part entière. Héritière de la philologie antique et médiévale, la philologie classique ne s'est intéressée aux « manuscrits » que pour l'établissement des textes, non pour étudier leur genèse, et les apparats critiques des éditions « savantes » n'ont jamais rien fourni d'autre que les quelques « variantes » (lexicales ou grammaticales) des dernières étapes génétiques (manuscrit au net, copie

[1] J'entends par « manuscrits » non pas le manuscrit « au net », la copie d'impression, mais les manuscrits *de travail* d'un écrivain – autrement dit : les brouillons.

d'impression, épreuves corrigées)[1]. Les manuscrits ne se laissant appréhender que sur un mode de lecture rhizomatique (et non linéaire), il fallait l'émergence d'un nouveau médium éditorial – l'édition hypertextuelle – pour rendre envisageable leur édition. La constitution de ces nouveaux corpus numériques est indissociable d'une réflexion approfondie sur les modes d'édition et d'annotation nécessaires pour rendre ces textes lisibles sans les dénaturer[2].
Le cas qui nous retiendra ici, l'édition de fonds épistolaires, est également un de ces cas où les possibilités offertes par les outils numériques transforment radicalement l'approche du matériau : longtemps considérées comme fonds de documents historiques (traces biographiques, documents d' « accompagnement » de la genèse des œuvres), les correspondances peuvent désormais constituer des corpus textuels, si un mode éditorial spécifique parvient à désancrer les lettres de leur concaténation chronologique – donc de la logique historico-biographique des éditions papier qui, jusqu'à présent, a imposé ce mode de lecture essentiellement documentaire des correspondances.

1.2. Le mode d'édition / d'annotation détermine le mode de lecture / d'usage des textes
Ce principe, qui vaut en règle générale, s'avère particulièrement crucial pour les ensembles fragmentaires que sont les correspondances. Tandis que les romans, les recueils de poèmes, les pièces de théâtre, les autobiographies, sont des textes constitués par l'auteur lui-même en volumes recevant leur unité et leur légitimité de ce geste éditorial (destiné au public), les correspondances, qu'elles soient monographiques (Proust-Gallimard) ou générales (*Correspondance* de Marcel Proust), sont des artefacts éditoriaux. Outre les questions pratiques posées par la notion de « recueil » de lettres (voir section 2 ci-après), se pose un problème fondamental de « lisibilité » qui tient au type même de discours qui définit le texte épistolaire : les lettres, n'ayant pas été écrites en vue de leur publication, ne produisent pas leur propre système de référents (ne sont donc pas auto-suffisantes, comme peut l'être un poème) mais s'articulent allusivement à des conversations privées et à des référents communs à l'épistolier et à son destinataire, qui échappent à un lecteur extérieur à cet échange (et en particulier au lecteur moderne, qui ne baigne pas dans l'univers de références historiques, politiques, culturelles du moment). La « lisibilité » de la lettre dépend donc étroitement de son co-texte éditorial, qui conditionne la réception du lecteur :
-dans le cas (rare) d'une relation épistolaire prolongée entre deux correspondants, et conservée dans sa (quasi-)totalité, une publication monographique (ainsi, Proust-Gallimard) offre un « texte » suivi, un *texte dialogique* qui se présente à la réception du lecteur comme l'histoire d'une relation (amicale, conflictuelle, etc.) entre un écrivain et son éditeur. L'annotation servant à préciser quelques références, bien que philologiquement indispensable, est à peine nécessaire au lecteur, pris dans la dynamique de l'échange ;
-dans les cas beaucoup plus fréquents où le destinataire n'a pas conservé ses lettres et ne publie que celles de Proust, la « lisibilité » de cet échange tronqué (par exemple, les lettres de Proust à Lucien Daudet – voir Daudet, 1929) se heurte fréquemment à l'écueil de la référence. Quelques annotations fournies par le destinataire lors de la première publication de ces lettres permettent de reconstituer l'objet de la discussion, et/ou le contexte. Mais comment lire ces lettres isolées de l'échange qui leur donne pleinement sens ? Daudet ayant, en outre, censuré à la publication une grande partie des passages concernant les médisances mondaines, les confidences ayant trait à la vie privée, les reproches ou les effusions sentimentales, pour ne retenir que les discussions générales et surtout littéraires, le lecteur est amené à lire ces lettres (tronquées) comme un florilège de jugements de Proust (sur la création littéraire, sur son œuvre, sur un certain nombre de

[1] Même lorsque des éditions modernes entendent donner un aperçu des brouillons de l'œuvre, il s'agit d'extraits décontextualisés, isolés des passages hétérogènes qui les précèdent ou qui les suivent dans le brouillon, isolés également des étapes rédactionnelles qui les précèdent ou les suivent dans l'élaboration de l'épisode concerné, et surtout, d'extraits rendus lisibles (ratures supprimées, etc.). Voir, par exemple, les « esquisses » fournies à la fin de chaque volume d'*À la recherche du temps perdu*, édition dirigée par J.-Y. Tadié, Gallimard, « Bibliothèque de la Pléiade », 1987-1989.
[2] L'HyperNietzsche (www.hypernietzsche.org/) constitue un exemple intéressant d'édition numérique de manuscrits fondée sur une philologie génétique adaptée à ce nouveau médium : à la fois archive virtuelle, corpus textuel (transcriptions établissant soigneusement le texte, avec ses ratures, ses additions interlinéaires, etc.), annotation savante, et guidage rhizomatique à travers les divers documents du corpus (chemins génétiques, chemins thématiques).

sujets d'intérêt général) – donc : comme un prolongement thématique ou métadiscursif du roman proustien ;
-cependant, dans la plupart des cas, les éditions scientifiques de correspondances se font sous forme de correspondances générales classées chronologiquement – et c'est aussi le cas de la correspondance de Proust : l'édition de monographies étant le fait des destinataires eux-mêmes dans les années qui suivirent la mort de l'écrivain, ces recueils à petits tirages sont devenus introuvables, et leurs lettres ont été reprises et intégrées dans la monumentale *Correspondance* en vingt et un volumes, classés chronologiquement[1]. Restituées à leur contexte historique et culturel par une annotation érudite qui identifie chacune des références littéraires, artistiques, politiques, mondaines, les lettres gagnent en signification et s'éclairent mutuellement (Proust tenant souvent des discours similaires à plusieurs correspondants auxquels il écrit le même jour). Mais ce mode d'édition, qui les classe chronologiquement, tend à faire des lettres les fragments d'un journal intime, des témoignages biographiques. Il induit une lecture documentaire, où la dimension éminemment relationnelle du texte épistolaire se perd.

1.3. L'épistolaire : entre document historico-biographique et classe textuelle spécifique

Pendant plusieurs décennies, peut-être parce que les écritures de l'intime étaient tenues en suspicion, les correspondances ont été tenues à l'écart du champ « littéraire », et tout juste considérées comme des documents à valeur essentiellement historique – pour l'établissement de biographies (genre lui-même déconsidéré). Depuis une ou deux décennies, par un prévisible mouvement de balancier, la revalorisation de l'intime a produit un intérêt spécifique pour l'épistolaire, considéré désormais comme genre littéraire (Kaufmann, 1990)[2]. Genre littéraire ? ou classe discursive particulière ?
Avant de se hâter de trancher la question, il convient de remarquer l'hétérogénéité du corpus, même dans une correspondance d'écrivain, et même d'un écrivain de la qualité de Proust. À partir de quel moment une lettre est-elle « littéraire » ? Une partie des lettres sont des billets relativement brefs et utilitaires, et la majorité des autres, même longues, discutent de questions pratiques, substituts de conversations téléphoniques[3] ; à une époque où il y avait plusieurs distributions de courrier par jour, et même des « petits bleus » (courriers par réseau pneumatique) qui arrivaient chez le destinataire en quelques minutes, une grande partie des lettres de Proust, « écrites au galop », ont la même fonction que nos courriers électroniques : demandes, réponses, poursuite d'une conversation orale, etc., elles ont une fonction relationnelle privée – et ne sont pas destinées, à l'inverse des lettres de la marquise de Sévigné ou de Bayle, à être lues dans les salons. En revanche, il est indéniable que ces objets textuels se caractérisent par une forme spécifique : en-tête et date (parfois), formule d'adresse toujours, formule d'adieu, signature, qui les classent comme « lettres » au premier regard, avant même la lecture du texte proprement dit. Inscription de la relation, et souvent des coordonnées du réel, ces indices formels, ainsi que la modalité généralement discursive, font de la lettre une parole *adressée*. Je propose donc de définir l'épistolaire simplement comme « classe de textes » spécifique (plutôt que d'un « genre littéraire »).
Afin que cette classe textuelle au fonctionnement particulier (texte écrit, mais doté d'un grand nombre de traits discursifs de l'oral) puisse être étudiée pour elle-même, il est nécessaire de constituer un corpus numérique. La vaste correspondance de Proust (plus de 6000 lettres, sans compter les lettres inédites réapparaissant occasionnellement) peut constituer un premier ensemble. La base gagnerait à être augmentée par d'autres correspondances de la même époque, permettant d'étudier régularités ou différences, et de s'interroger sur la notion de « style épistolaire ».
Il n'en reste pas moins que ces textes à forte composante discursive, sont fondamentalement inscrits dans un contexte historico-culturel qui est leur objet premier de référence, et que la richesse de l'expérience culturelle de Proust fait de sa correspondance une base de documentaire importante pour l'étude de la culture de la Belle-Époque et de la première guerre mondiale, ce qui

[1] Marcel Proust, *Correspondance*, texte établi, présenté et annoté par Philip Kolb, Plon, 1970-1993, 21 tomes.
[2] L'Association Interdisciplinaire de Recherche sur l'Épistolaire (A.I.R.E.) publie une revue semestrielle (*Bulletin de l'AIRE*, puis *Revue de l'AIRE*) depuis 1988.
[3] Proust a eu le téléphone très tardivement chez lui, l'utilisait peu, laissant ses domestiques répondre à sa place et communiquer ses messages, et il finit par s'en passer totalement pour ne pas être dérangé.

constitue actuellement la principale motivation des demandes sociales pour la constitution d'une telle base numérique.

1.4. De l'annotation traditionnelle (éclaircissements référentiels) à une annotation multi-critères, en vue de l'extraction de sous-corpus de recherche et d'exploitations statistiques

La constitution d'une base « correspondance de Proust » se trouve ainsi au centre d'une double demande et d'un triple enjeu :

- d'abord, une forte demande sociale (et non pas uniquement des spécialistes de Proust) de disposer du texte numérisé de ces six mille lettres (ou 21 volumes) afin de faciliter les recherches en plein texte ;
- corrélativement, une demande d'annotation précise des références, qui permette de saisir les allusions de l'épistolier ; donc, implicitement, d'une base documentaire
- mais parce qu'une lettre n'est pas uniquement un objet documentaire historique, il serait judicieux de profiter de la constitution de cette base numérique pour proposer une annotation (un balisage) qui permette aussi, par la suite, de mener des recherches sur le « genre » épistolaire en tant que tel, et d'étudier, dans cette perspective, le « style » épistolaire proustien.

Comment concilier ces demandes et ces enjeux ?

L'annotation de Philip Kolb étant extrêmement érudite et précise, il n'est pas difficile –et il est nécessaire- de l'introduire sous forme de notes et de liens hypertextuels, pour fournir au tiers-lecteur les éléments contextuels de référence. Mais permettre des recherches sur l'épistolaire proprement dit requiert de nouveaux types d'annotations. Lesquels ?

La question peut se reformuler ainsi : que peut-on vouloir chercher dans un corpus épistolaire ? comment l'utilisateur peut-il souhaiter interroger ce corpus ? Ces recherches étant encore inédites (la masse des 6000 lettres est ingérable sous sa forme papier actuelle, même pour les spécialistes…), il faut donc *imaginer* les requêtes possibles sur un corpus épistolaire numérique…

- Une partie des requêtes sera sans doute d'ordre thématique (retrouver aisément toutes les lettres traitant de politique, ou de musique, ou de littérature, ou du roman proustien, entre autres possibilités[1]..). Faut-il intégrer un logiciel de traitement de corpus, afin que par des requêtes lexicales (listes de mots) l'utilisateur repère des occurrences thématiques ? On examinera cette question ci-après (3.2).

- Une étude thématique ne peut être productive que si elle est distributionnelle (avec quel correspondant tel sujet est-il plus fréquemment abordé ?)

- Une étude centrée sur le style épistolaire devrait pouvoir étudier les formules d'adresse, de politesse, d'adieu, les citations, et les comparer (analyse distributionnelle par correspondants, ou par périodes). Ce qui requiert un balisage de ces éléments de structure. (Voir 3.1)

-Par ailleurs une lettre, parole « adressée », est toujours destinée à agir sur le destinataire, qu'il s'agisse de lui demander un service, d'influencer ses décisions, de s'excuser ou de se justifier d'une accusation, ou au contraire l'amuser, de maintenir la relation par un échange de nouvelles, par la sollicitude : la dimension « expressive » de l'épistolier est subordonnée à une visée pragmatique. Quel genre d'épistolier est Proust ? On peut vouloir mener ces études statistiquement à travers l'ensemble du corpus, mais aussi de façon distributionnelle (étudier le type de relation entretenu avec tel ou tel correspondant). Seul un balisage des opérations pragmatiques permettrait ces études spécifiques au style épistolaire. (Voir 3.3)

- Toutefois, le corpus épistolaire, je l'ai souligné plus haut, n'est pas homogène. Il est constitué de genres (ou sous-genres) conventionnels : lettre de condoléances, de félicitations, accusé de réception, ordre de ventes à la Bourse, lettre de vœux, dédicaces, lettres administratives, ces genres étant relativement codifiés – même si l'épistolier peut introduire des écarts par rapport aux normes sociales ; en revanche, d'autres genres épistolaires sont moins conventionnels, telles les nouvelles que l'on échange, les échanges de vues amicaux, les négociations avec l'éditeur. Les genres très codifiés imposant une attitude pragmatique particulière, l'usager peut vouloir exclure du corpus certains types de lettres, pour ne pas générer de bruit ou ne pas fausser les résultats dans certaines requêtes ; ou au contraire, il peut souhaiter

[1] Actuellement, l'index des noms de personnes ne permet que de retrouver des passages où il est question de tel ou tel homme politique, musicien, etc., ce qui est très limitatif.

n'étudier que certains types très codifiés, pour voir si Proust y adopte le comportement socialement attendu…. (et plus tard, si la base parvient à intégrer plusieurs correspondances, pour mener des études comparatives – rhétoriques, socio-linguistiques, etc. – sur certains genres épistolaires codifiés).
Il faut donc un balisage multiple qui permette de sélectionner ou d'exclure dans l'ensemble du corpus des sous-corpus (par dates, par correspondants, par thèmes, par type de lettre), et qui permette d'interroger le corpus choisi en fonction d'un ou de plusieurs critères croisés.

2. Rapide examen critique de corpus épistolaires numérisés ou numériques

L'apparition de la TEI et du langage XML, qui décuplent les possibilités d'annotation et rendent pensable la production d'éditions web-centrées facilement accessibles, est très récente, et le caractère expérimental, dispersé, de leurs applications rend difficile une vue globale sur les corpus numériques en cours de constitution, notamment dans le domaine encore peu étudié de l'épistolaire.
L'examen critique de deux bases très différentes dans leurs objectifs comme dans leurs pratiques annotatives, Frantext, et Arcane, me fournira quelques remarques méthodologiques sur les écueils et les limites de certaines pratiques éditoriales numériques.
L'objectif de Frantext étant, comme chacun sait, de constituer une base de données lexicales pour l'établissement du *Trésor de la Langue française*, les textes qui y sont intégrés, quel que soit leur genre, sont étiquetés uniquement selon une perspective lexicale et morpho-syntaxique (afin de permettre la recherche de fréquences lexicales ou de constructions syntaxiques, de voisinages et de co-occurrences). Un maniement adroit des listes de mots permet des repérages thématiques dans les textes (à condition que l'expression de l'auteur ne soit pas trop métaphorique ou allusive : voir 3.2). Mais outre la difficulté de ces requêtes pour un usager non expert, et l'impossibilité fréquente de visualiser les occurrences dans des contextes élargis, la philosophie éditoriale de Frantext ne peut convenir à l'édition de lettres : comme le montre une recherche effectuée sur la correspondance de Flaubert, l'unité textuelle, pour Frantext, est l'item bibliographique, c'est-à-dire chaque tome de l'édition imprimée numérisée (ici, l'édition Conard des années 1920). À la différence des recueils de poèmes définis (selon un groupement esthétique ou chronologique) par l'auteur lui-même, les tomes d'une correspondance générale recueillie et éditée à titre posthume sont des artefacts éditoriaux : ils regroupent parfois quatre ou cinq années, parfois une année, de la correspondance de l'écrivain, pour des raisons aléatoires (il faut un certain nombre de lettres pour faire un volume éditorialement viable). On peut étudier dans Frantext si tel mot (« roman » par exemple) est plus souvent employé dans le tome 1 que dans le tome 2 ou dans le tome 12, mais quelle est la valeur de ces informations, dès lors que le découpage des lettres en tomes est aléatoire, et que les lettres de 1844 qui appartiennent au tome 1 pourraient tout aussi bien se trouver dans le tome 2 si Conard avait décidé de faire des volumes moins épais ? Outre que le séquençage du corpus par années réelles est impossible, une analyse distributionnelle par correspondants est également impossible. Numériser une édition papier existante et la considérer comme « un » texte nie la *spécificité de la lettre comme unité textuelle*. Même s'il est tentant de numériser au kilomètre une édition imprimée (c'est rapide, économique), il convient donc de refuser cette solution.
La correspondance de Bayle, en cours d'édition par Anthony McKenna[1], propose au contraire un classement par lettre, chacune étant définie par son destinateur, son destinataire, sa date, son lieu d'expédition et de destination. L'utilisateur peut donc choisir de faire des recherches à travers l'ensemble du corpus, ou de sélectionner des sous-corpus (par épistolier, par destinataire, par date ou tranche chronologique, par lieu). Cette reconfiguration possible du corpus à chaque requête tient compte de la spécificité du fonds. Le balisage (par hyperliens de couleurs différentes) offre des annotations philologiques (ratures, additions, etc.) et des annotations référentielles (identification de personnes, d'œuvres). Un logiciel de cartographie permet de visualiser et de mesurer statistiquement les échanges épistolaires entre les correspondants, répartis en plusieurs points de l'Europe. L'objectif de cette base est documentaire et culturel : il s'agit d'étudier la diffusion de la pensée des Lumières à travers l'Europe du XVIIIe siècle. Ces lettres, en général fort longues, soigneusement écrites, pleines d'informations savantes et de discussions scientifiques et

[1] Édition numérique dans une base de données Arcane développée par Éric-Olivier Lochard à l'Université Jean Monnet de Saint-Étienne ; et, parallèlement, édition imprimée, chronologique et annotée, à la Oxford Foundation.

philosophiques, sont en effet destinées majoritairement à être lues devant un public (parfois familial, souvent un cercle savant). Cependant, même si la relation intersubjective et les visées pragmatiques y tiennent moins de place que le contenu informatique (documentaire), ces traits spécifiquement épistolaires ne sauraient être absents de ces lettres – mais aucun balisage thématique ni formel n'est prévu, à ma connaissance, non plus qu'aucun outil permettant l'étude du lexique ou du style des épistoliers.
Comme pour l'édition imprimée (voir 1.2), les présupposés scientifiques qui ont présidé à la constitution des bases numériques induisent la détermination de leurs outils de recherche et donc une certaine limitation dans l'utilisation de la base, au détriment d'autres questionnements auxquels ces textes pourraient, légitimement, être soumis.

3. Pour un balisage spécifique des corpus épistolaires modernes : illustration

Les réflexions qui suivent s'appuieront sur quelques exemples tirés de la correspondance de Proust (voir Annexe 1, exemples 1 à 3), pour interroger la pertinence des outils existants et proposer – de façon plus empirique que théorique – des modes de balisage spécifiques pour une étude non seulement historico-documentaire, mais aussi générique et culturelle de l'épistolaire.

3.1. Balisage formel ou balises de structure (constituants formels de la lettre)

Dès lors que le balisage XML autorise une identification multiple d'un « document » – les balises « auteur », « destinataire », « lieu », « date », étant considérées comme balises de « structure » –, on peut étendre la nomenclature de ces balises de structure à l'étiquetage d'autres constituants génériques comme les formule d'adresse et d'adieu, la signature, les Post-scriptum. Un autre trait formel participe de la poly-énonciation épistolaire : les citations. Notre approche se veut ici purement descriptive : repérer les éléments formels entrant dans le texte de la lettre, qu'ils soient spécifiques de l'épistolaire ou également présents dans d'autres formes de texte. (Ainsi, le statut de la citation épistolaire gagnerait à être discuté d'un point de vue théorique : il y a plusieurs sortes de citations, et qui remplissent plusieurs fonctions ! Étudier si le régime citationnel des lettres en regard du régime citationnel romanesque requerrait un examen comparatif de ces deux types de corpus, ce qui supposerait des corpus numériques étiquetés de sorte à faire apparaître les citations. Au stade où nous en sommes, il est donc prématuré de décider si tel usage de la citation est littéraire et tel autre spécifiquement épistolaire. Toutefois, on peut remarquer que l'échange épistolaire introduit une dimension intertextuelle spécifique : la citation ou la reformulation des paroles du destinataire auxquelles l'épistolier répond – trait spécifiquement dialogique).

Exemple de balisage auteur/destinataire :

[MARCEL PROUST À ANTOINETTE FAURE] : [<name type=auteur> MARCEL PROUST</nom> À <nom type=destinataire> ANTOINETTE FAURE</nom>]

Exemple de balisage (simplifié) des formules d'adresse et de politesse :

Ma chère Antoinette : <adresse> ma chère <nom> Antoinette</nom>, </adresse>
Mon cher petit grand'père : <adresse> Mon cher petit <nom>grand'père</nom>,</adresse>

Le balisage est ici simplifié pour les besoins de l'exposé. Il va de soi, notamment, que la balise « nom » doit préciser qu'Antoinette est Antoinette Faure ou que le grand-père est Nathé Weil, afin qu'une recherche des occurrences de « Nathé Weil » puisse sélectionner cette occurrence.

Exemple de citations qui doivent être balisées (afin de permettre des recherches sur l'univers intertextuel de Proust aussi bien que des recherches sur la polyphonie énonciative) :

soldat « simple et sublime » comme dit le Petit Boulangiste
« Gais et contents nous allions triomphants »
« C'est Boulange, lange, lange, »

Dans la lettre à Halévy (Annexe 1, exemple 3), la question se pose de savoir s'il faut baliser comme « citation » l'adjectif «décadent » souligné par l'auteur : « Je ne suis pas *décadent* » suppose que le destinataire a traité le jeune Proust de décadent, et qu'il s'agit là d'une

reformulation allusive. Cet exemple montre la différence entre une description simple des typologies textuelles (texte citationnel vs texte de l'auteur) et une approche pragmatique des phénomènes discursifs, qui ne se contente pas des marques formelles (guillemets) ou des signes explicites (« comme dit » X), mais qui interprète le texte épistolaire dans le contexte de la relation discursive des deux amis. Peut-être le recours à une balise « allusion » pourrait-il résoudre la différence entre ces deux types de pratique intertextuelle.

3.2. Balisage thématique

Les requêtes thématiques constituent évidemment l'un des types de recherche privilégiés par les usagers. Les logiciels de traitement de corpus, s'ils sont très simples, ne permettent pas beaucoup de résultats et génèrent beaucoup de bruit ; s'ils sont complexes (comme Frantext), ils requièrent une compétence qui décourage même des étudiants avancés. Guider l'usager dans le corpus nécessite un balisage thématique.

En tant qu'éditrice (putative) d'un corpus épistolaire, je souhaiterais pouvoir disposer d'outils de balisage sémantique automatisé, mais ces outils sont-ils fiables sur des textes qui, comme ceux de Proust, du fait de leur style allusif et métaphorique, nécessitent une large part d'interprétation ?

Les trois lettres de l'Annexe montrent des pratiques sémantiques hétérogènes : dans la lettre 2, les substantifs « somme » et « francs » (bien qu'ambigus pris chacun isolément) peuvent être, du fait de leur collocation, automatiquement étiquetés dans une nomenclature de type « argent » ; quant à « bordel », etc., il mène sans ambiguïté à un classement du passage dans « sexualité », par exemple. De même, dans la lettre 3, la redondance des termes « amour », « aimer », etc., « chair », « genoux », permet un classement automatique sous les rubriques d' « amour » (ou « sentiment ») et « sexualité ». – En revanche, au début de la lettre 3, seul le terme « poète » permet d'assigner mécaniquement l'étiquette « littérature » (ou « poésie ») à une discussion allusive qui multiplie les référents culturels. Et dans la lettre 1, aucun terme explicite ne permet de définir que la première séquence concerne la politique. En effet, « soldat » et « général » ont toute chance de produire un étiquetage dans la catégorie « guerre » (par exemple), de sorte que la lettre pourrait apparaître dans une requête concernant la première guerre mondiale, mais nullement dans une recherche de lettres relatives à la politique… Même un historien cherchant « Boulanger » ne pourrait tomber sur cette lettre, le nom de Boulanger ne figurant explicitement nulle part dans le texte. Ce n'est qu'un exemple parmi quantité d'autres, Proust étant maître dans l'art de l'allusion culturelle. Ne faut-il pas, dans ce cas, envisager un balisage manuel du corpus ? (C'est une question…)

3.3. Balisage des opérations pragmatiques

La correspondance étant une pratique essentiellement interactive, relationnelle, le « style », la « rhétorique » du texte épistolaire ne peuvent s'appréhender, me semble-t-il, en dehors du cadre théorique de la linguistique pragmatique. L'épistolier est-il prompt à la récrimination, plaintif, toujours porté à demander des services, ou au contraire prompt à faire rire son correspondant ? ou a-t-il une tendance marquée à la confidence, à l'effusion lyrique, à l'expression de ses doutes et de ses craintes ? Seul un balisage spécifique permettra de mesurer objectivement, statistiquement, le « style » épistolaire, globalement ou de façon distributionnelle. Mais une fois énoncée cette évidence, comment opérer un balisage des opérations pragmatiques ?

Il faudrait pouvoir disposer d'une nomenclature stable, scientifiquement reconnue et acceptée, des « actes de langage ». Là réside un certain nombre de difficultés. Austin (1962) distingue les actes illocutoires par lesquels le locuteur « fait » quelque chose et les actes perlocutoires par lesquels le locuteur vise un certain effet. Tout d'abord, faut-il baliser les actes illocutoires, ou les actes perlocutoires (le but visé, la finalité) ? Mais la taxinomie des actes illocutoires est elle-même discutable. Ainsi, Austin divise ces actes illocutoires en cinq catégories : les verdictifs (exercice d'un jugement : décréter que, estimer, coter, juger, etc.) ; les exercitifs (exercice d'un pouvoir institutionnel de la parole : ordonner, excommunier, nommer, renvoyer, etc.,) ; les promissifs (promettre, jurer de, donner sa parole, etc.) ; les comportatifs (catégorie assez disparate, de tout ce qui a trait au comportement social : remercier, s'excuser, compatir, critiquer, applaudir, etc.) ; et enfin les expositifs (catégorie « difficile à définir » de l'aveu même du théoricien, regroupant l'argumentation, l'explication : dire, affirmer, nier, concéder, etc.). Austin reconnaît cependant que sa classification est inachevée et imparfaite. – Si, pour la lettre 3 (Annexe) la catégorie des expositifs est satisfaisante pour la plupart des séquences (Proust expose sa façon de voir,

argumente), dans le cas de la lettre 2, où ranger la « réclamation » ? La justification (il est impossible d'échouer deux fois) relève de l'argumentation, certes. Mais la « demande » initialement formulée est un acte de langage spécifique, antérieur à toute argumentation, et doit pouvoir être catégorisé comme tel. En outre, que faire du micro-récit des déboires du jeune homme ? Selon Austin, un récit relève de la catégorie du constat, qu'il tient en dehors des actes illocutoires (actes effectués *en* disant quelque chose, par opposition à l'acte *de* dire quelque chose » -Austin, 1962, 113, Eluerd, 150). Pourtant, en racontant ses déboires, le jeune homme *fait* deux choses : il fait un aveu, il justifie sa demande d'argent : le récit n'est donc pas simplement un énoncé constatif. En outre, sa visée perlocutoire est claire : apitoyer son aïeul, pour obtenir la somme recherchée. En revanche, dans la lettre 1, le récit (les différentes séquences narratives) n'ont aucune valeur illocutoire particulière dans la classification d'Austin, mais cet égrenage de petits faits quotidiens « fait » quelque chose (donner des nouvelles) et a une fonction particulière : maintenir la relation entre amis pendant les vacances…

Une autre classification ne pourrait-elle pas mieux rendre compte de ces actes de langage ? Searle, critiquant Austin, propose quelques années après une taxinomie fondée non seulement sur le but illocutoire, mais aussi sur l'état d'esprit du locuteur et sur le rapport de son discours au monde. Cinq catégories également : les assertifs (affirmations, assertions, descriptions, caractérisation, explications, etc.) ; les directifs (essayer de faire faire quelque chose par l'interlocuteur : demande, ordonner, supplier, interdire, etc.) ; les promissifs (s'engager à faire quelque chose) ; les expressifs (exprimer un état psychologique : s'excuser, féliciter, remercier, etc.) ; les déclarations (provoquer un changement par une déclaration ; il s'agit de performatifs : démissionner, déclarer la guerre, excommunier, etc.). Cette taxinomie souple, qui peut paraître séduisante, a été très sévèrement critiquée par les théoriciens de la linguistique pragmatique, qui lui reprochent d'ignorer les usages ordinaires de la langue (voir Eluerd, p.165-175) : elle reproduit, en fait, les fonctions de la communication de Jakobson (fonctions référentielle, conative, expressive, etc.), est centrée sur le locuteur exclusivement, et ne tient compte ni du contexte, de la situation de communication, ni de l'interlocuteur.

Si dans une conversation orale les locuteurs sont, en effet, des co-locuteurs qui construisent ensemble la modalité de l'échange et son orientation, dans le cas des lettres, cependant, nous avons chaque fois affaire à une énonciation singulière (même si celle-ci est ouverte en permanence sur l'univers discursif des deux correspondants qui affleure en permanence dans le texte épistolaire).

Sans vouloir adopter la taxinomie de Searle, contestable, et en tenant compte de la visée implicite de tout « acte de langage » épistolaire, il me semble qu'il faudrait trouver, empiriquement, un étiquetage qui tienne compte à la fois de ce que « fait » la lettre et ce qu'elle « vise ».

Dans la lettre 1 (Annexe), Proust raconte : il donne des nouvelles (c'est ce qu'il fait), la visée étant (possiblement) de maintenir le contact, de partager une communauté d'opinion. Dans la lettre 2, en revanche, quand il raconte, il fait un aveu, et cet aveu a apparemment pour but de justifier sa demande (« voici pourquoi » : rhétorique de justification), mais en réalité d'apitoyer le grand-père. Un même « type » textuel, le récit, a dans ces deux *contextes* une fonction et une finalité différentes.

On pourrait donc imaginer une classification assez large de ce que « fait » l'épistolier (donne des nouvelles, confidence/aveu, exprime ses idées, polémique, donne une invitation/un rendez-vous, refuse une invitation, etc.), qui, couplée à une nomenclature thématique (scolarité, politique, amour, argent, littérature, etc.), permettrait des recherches relativement précises à travers le corpus.

Faut-il en rester là, et laisser à l'usager le soin d'approfondir l'analyse pragmatique à partir de sous-corpus qu'il aura sélectionnés ? C'est une option. Pourtant, il me semble qu'il serait intéressant de baliser aussi la *visée* des énoncés épistolaires. Ainsi, dans la lettre 2, la maxime (« il n'arrive pas deux fois dans la vie ».. etc.) a une finalité justificative ; de même, dans la lettre 3, les opinions littéraires émises par le jeune homme visent à le justifier d'accusations portées contre lui. Même chose pour la phrase finale, sur l'équivalence éthique de toutes les formes d'amour. Ce souci de justification me paraît récurrent dans la correspondance de Proust, et gagnerait à être quantifié.

Conclusion

Ce programme d'édition et de balisage, quoique copieux, peut paraître insuffisant dans ses ambitions linguistiques, ne prévoyant pas d'inclure certains outils d'analyse lexicale, grammaticale, prosodique, etc. existant dans d'autres corpus numériques.

Reste à savoir si l'aide à la lecture doit viser l'exhaustivité. La base doit-elle permettre de tout faire ? Il me semble que dans un vaste corpus épistolaire, défini par sa double appartenance documentaire et textuelle, l'aide à la lecture consiste à fournir au lecteur des outils de repérage à travers l'hétérogénéité des lettres, afin qu'il puisse extraire les passages qui correspondent au type de recherches spécifiques à l'épistolaire qu'il est susceptible de vouloir mener dans ce corpus (recherches sur des genres épistolaires précis : lettres d'affaires, de condoléances, etc. ; recherches sur des catégories thématiques ou des pratiques culturelles : lettres sur la musique, citations, etc. ; recherches sur les interactions épistolaires : demandes de service, récriminations, attitude défensive, etc.) – ensuite, une fois extraits tous les passages pertinents, l'usager ne pourra-t-il pas en étudier le style, le lexique, les caractéristiques énonciatives par ses propres moyens, ou les soumettre à d'autres outils d'analyse ?

Mais la discussion reste ouverte. Ce sont des propositions à caractère programmatique qui peuvent encore évoluer…

BIBLIOGRAPHIE

AUSTIN, J. L. 1962. *How to do things with words*, Cambridge, Mass.: Harvard University Press, (*Quand dire c'est faire*. Introduction et traduction de G. Lane. Paris, Le Seuil, 1970).

DAUDET, L. 1929. *Autour de soixante lettres de Marcel Proust*. Paris, Gallimard, *Les Cahiers Marcel Proust*, n° 5.

ÉLUERD, R. 1985. *La Pragmatique linguistique*, Paris, Nathan.

KAUFFMANN, V. 1990. *L'Équivoque épistolaire*, Paris, Minuit.

KERBRAT-ORECCHIONI, C. 1990, 1992, 1994. *Les Interactions verbales*, Paris, Armand Colin, 3 tomes.

PROUST, M. *À la recherche du temps perdu*, Édition publiée sous la direction de Jean-Yves Tadié, Paris, Gallimard, 1987-1989, Bibliothèque de la Pléiade, 4 tomes.

PROUST, M. *Correspondance*, Texte établi, présenté et annoté par Philip Kolb, Paris, Plon, 1970-1993, 21 tomes.

PROUST, M. GALLIMARD, G. *Correspondance 1912-1922*, Édition établie, présentée et annotée par Pascal Fouché, Paris, Gallimard, 1989.

RASTIER, F. 2001. *Arts et sciences du texte*, Paris, Presses universitaires de France.

SEARL, J. R. 1982. *Sens et expression. Études de théorie des actes de langage*, Traduction et préface de J. Proust, Paris, Minuit, (*Expression and Meaning*. New-York, 1979).

ANNEXE
Exemple 1 :

[MARCEL PROUST À ANTOINETTE FAURE]

[Paris le 15 juillet 1888]

Ma chère Antoinette,

Croiriez-vous que Maman m'a déchiré une lettre pour vous. L'écriture était trop mauvaise. Au fond je crois qu'un grand éloge de notre brave général, du soldat « simple et sublime » comme dit le Petit Boulangiste a excité les vieux sentiments orléanistes-républicains de madame Jeanne Proust.**[N1]**
Jamais les rues d'Auteuil (où j'ai passé seulement la journée du 14) n'avaient été aussi animées qu'hier. Vous ne trouvez pas entraînant ce refrain :
« Gais et contents nous allions triomphants »
ou :
« C'est Boulange, lange, lange, »
hurlé par tous, femmes, ouvriers, jusqu'aux petits enfants de cinq à huit ans qui le chantent très très juste –avec ardeur.
Quoique l'homme soit très commun et un vulgaire batteur de grosse caisse, ce grand enthousiasme si imprévu, si *roman* dans la vie banale et toujours la même, remue dans le cœur tout ce qu'il y a de primitif, d'indompté, de belliqueux. **[N2]**
Vous voyez que je ne suis pas grand philosophe et je ne trouve guère que des adjectifs quand je cherche des raisons qui (pardonnez-moi cette enfilade de qui) me donnent envie de brailler : Il reviendra. **[Métadiscours]**
Je n'ai rien à vous raconter des Champs-Élysées. Blanche est toujours très douce, d'un visage angélique espiègle et résigné. Marie Bénardaky est très jolie et de plus en plus exubérante. Elle s'est *battue* à coups de poing avec Blanche qui a été *battue* et qui (ceci n'a pas de rapport) vous fait beaucoup remercier de votre lettre. **[N3]**
J'ai composé avant-hier au concours cinq heures de suite sans l'ombre d'un repos. Je suis arrivé à la Sorbonne à 9 h. 30 et j'en ai quitté à 4 heures moins un quart. La composition a duré de 10 h. 30 à 3 h. 30. Nous étions 120 ou 130 composants, c'est-à-dire les deux (rarement les trois) premiers de toutes les divisions de tous les lycées. C'était en histoire. **[N4]** J'irai entendre Paulus un de ces soirs. Je vous rendrai compte de la représentation. **[N5]**
Faut-il dire quelque chose de votre part à vos amies ? Je vais à peu près tous les jours aux Champs-Élysées. **[Offre de services, motivée]**
Je vous souhaite de bonnes et charmantes vacances ainsi qu'à toutes celles de vos amies que je connais.
Présentez mes affectueux respects à Monsieur et Madame Faure, à Mademoiselle Lucie, Mademoiselle Marcelle et sa sœur etc.

Marcel Proust.

N1-N2 : nouvelles personnelles [N1] et collectives [N2]. **Thèmes** : la politique, le boulangisme
N3 : nouvelles collectives (cercle restreint). **Thème** : anecdotes sur des amies communes
N4 : nouvelles personnelles. **Thème** : le concours général, la scolarité de Proust
N5 : nouvelles personnelles (projet). **Thème :** le café-concert, les arts du spectacle
Ces séquences sont coupées par des remarques de l'épistolier sur son style **[métadiscours]** et se terminent par une **[offre de service]**, suivie des formules d'adieu d'usage dans un texte de type épistolaire.

Exemple 2 :

[MARCEL PROUST À NATHÉ WEIL]

Jeudi soir [17 mai 1888]

Mon cher petit grand'père,

Je viens réclamer de ta gentillesse la somme de 13 francs que je voulais demander à Monsieur Nathan, mais que Maman préfère que je te demande. **[D]** Voici pourquoi. J'avais si besoin de voir une femme pour cesser mes mauvaises habitudes de masturbation que papa m'a donné dix francs pour aller au bordel. Mais 1° dans mon émotion j'ai cassé un vase de nuit, 3 francs 2° dans cette même émotion je n'ai pas pu baiser. Me voilà donc comme devant attendant à chaque heure davantage 10 francs pour me vider et en plus ces 3 francs de vase. Mais je n'ose pas redemander sitôt de l'argent à papa **[récit, à valeur explicative]** et j'ai espéré que tu voudrais bien venir à mon secours dans cette circonstance **[D']** qui tu le sais est non seulement exceptionnelle mais encore *unique* : il n'arrive pas deux fois dans la vie d'être trop troublé pour pouvoir baiser [.] **[maxime à finalité justificative]**
Je t'embrasse mille fois et n'ose te remercier d'avance. **[R]**
Je passerai demain à onze heures chez toi. **[rendez-vous]** Si ma situation t'a ému et que tu te rendes à mes prières j'espère que je te trouverai ou un commissionnaire chargé de la somme. **[D'']** En tous cas merci car ta décision n'aura pour cause que ton amitié pour moi [.] **[R']**

Marcel

D = demande
D' = réitération de la demande
D'' = seconde réitération de la demande
Récit (ou confidence) : différence pragmatique avec N1, N2 etc. de la lettre 1. Ici, la narration a une valeur explicative, sert à motiver la demande D (On pourrait aussi bien intituler cette section : explication, ou motivation).
R : remerciement (on passe ici sur la litote)
R' : réitération du remerciement.

Exemple 3 :

[MARCEL PROUST À DANIEL HALÉVY]

[Avant le mardi 22 mai 1888]

Mon cher Daniel

On géographise avec zèle autour de moi. Je me donne deux minutes de répit –. Je ne suis pas *décadent*. Dans ce siècle j'aime surtout Musset, le père Hugo, Michelet, Renan, Sully Prud'homme [sic], Leconte de Lisle, Halévy, Taine, Becque, France. Je me plais beaucoup à Banville, à Hèrèdia [sic] et à une certaine anthologie <idéale> composée de morceaux exquis de **poètes** que je n'adopte pas en entier : La Création des Fleurs de Mallarmé, des Chansons de Paul Verlaine etc. etc. – Mais j'ai horreur des critiques qui ont une attitude ironique vis-à-vis des décadents. Je crois qu'il entre dans leur cas beaucoup d'insincérité, mais inconsciente ou au moins sans clairvoyance. Les causes de cette insincérité sont si tu veux, la religion des belles formes de langage, une perversion des sens, une sensibilité maladive qui trouve des jouissances très rares dans de lointaines accordances, dans des musiques plutôt suggérées que réellement existantes.
Quant au **style** Mendès, Silvestre, Banville (en **prose**) je crois qu'il mène à l'insincérité qui est le commencement de la banalité (ça ressemble un peu à M. Purgon : la dissenterie etc. etc.) : Si ça ne te semble pas très clair je t'expliquerai de vive voix. –
Je n'ai pas de ***passion*.** Je trouve ton ami le plus cher ou celui qui t'aime plus que les autres, je ne sais pas comment cela se dit en français – très gentil et j'ai un très réel plaisir – que je n'essaye pas de dissimuler – à me trouver avec lui. Mais comme j'en ai autant à me trouver avec toi, je voudrais que tu me dises quels jours tu ne rentres pas immédiatement à 4 heures. Et ceux où tu voudras bien de moi, je serai à tes ordres. Bizet trouvera que je commence avec toi la série des « listes » qui sont comme tu sais, chez moi le commencement de l'**amitié**. Il se tromperait absolument, car la mienne pour toi est déjà ancienne et quoique ce soit assez bête à dire mais enfin sur le papier tout passe, très vive.
Je te remercie de m'avoir donné cette occasion de ne pas écouter Choublier. D'ailleurs en t'écrivant je croyais te causer et je me suis ainsi donné l'illusion d'un grand plaisir, ce qui tu sais, n'est pas plus illusoire.

Post Scriptum

Je te propose de fonder avec moi (mais soyons seuls, directeurs) un grand **journal** d'**art**. – . Quant à ton **pédéraste** virtuel ou non, tu peux très bien te tromper. Je sais… qu'il y a des jeunes gens… (et si ça t'intéresse et que tu me promettes un *secret absolu*, même pour Bizet, je te donnerai des pièces d'un intérêt très grand à ce point de vue, à moi appartenant, à moi adressées) des jeunes gens et surtout des types de huit à dix-sept ans qui ***aiment*** d'autres types, veulent toujours les voir (comme moi, Bizet) pleurent et **souffrent** loin d'eux, et ne désirent qu'une chose les **embrasser** et se mettre sur leurs **genoux**, qui les **aiment** pour leur ***chair*,** qui les couvrent des yeux, qui les appellent **chéri**, mon ange, très sérieusement, qui leur écrivent des lettres **passionnées** et qui pour rien au monde ne feraient de la **pédérastie**.
Pourtant généralement l'**amour** l'emporte et ils se **masturbent** ensemble. Mais ne te moque pas d'eux et de celui dont tu me parles, s'il est ainsi. Ce sont en somme des **amoureux**. Et je ne vois pas pourquoi leur **amour** est plus malpropre que l'**amour** habituel.

DE L'UTILITÉ DE LA SÉMANTIQUE TEXTUELLE COMME MEDIUM ENTRE CORPUS ET ANALYSE

LES JEUNES DE CHIRAC ; ANALYSE SUR CORPUS NUMÉRISÉ

Baptiste FOULQUIÉ
CPST, Université de Toulouse 2

SOMMAIRE

Résumé : *Le développement des logiciels dit d'analyse de texte s'accompagne de nombreuses analyses sur grands corpus et intéresse la sémantique.*
Nous tenterons de voir dans ce travail, comment la sémantique peut tirer avantage des outils que sont ces logiciels, sans pour autant leur laisser la place qui est la sienne dans le champ de l'analyse du contenu des textes. Nous tenterons de faire la part, concernant l'analyse thématique, entre ce qui relève de la compétence du manipulateur du logiciel, et ce qui relève de la compétence du sémanticien. Nous nous plaçons ici dans la perspective de F.Rastier, et tentons d'utiliser les concepts de la sémantique interprétative (isotopies et paratopies, thèmes spécifiques et molécules sémiques).
Prenant prétexte des travaux de Damon Mayaffre, à qui nous devons notre corpus, nous verrons l'importance que peuvent prendre les précautions méthodologiques, notamment en ce qui concerne la répartition raisonnée des textes lors de la création du corpus, sur l'interprétation des textes.
Nous reviendrons ensuite sur les concepts qui justifient l'analyse des cooccurrences, en nous concentrant sur la paratopie qui permet l'établissement des molécules sémiques qui représentent les thèmes.

1. Préambule

La numérisation de grands corpus ainsi que l'apparition de logiciels permettant leur traitement permet aujourd'hui l'émergence d'une pratique analytique nouvelle : l'analyse textuelle assistée par ordinateur. Notre propos sera ici prioritairement centré sur la pratique qui consiste à passer un corpus au crible d'un logiciel (ici Hyperbase). En effet, l'ergonomie logicielle est telle qu'elle n'incite pas toujours l'utilisateur à avoir un regard critique sur sa pratique, mais elle peut en revanche l'inciter à dégrader sa théorie en fonction des limitations et des objectifs de la pratique. Cette adaptation étant nécessaire, il n'en reste pas moins qu'elle doit être questionnée et si possible limitée.
Cette étude prend prétexte des travaux réalisés par Damon Mayaffre dans son ouvrage : « Paroles de président, Jacques Chirac (1995-2003) et le discours présidentiel sous la V^{ème} république ». Rappelons que le corpus de Mayaffre est composé de l'intégralité des discours des présidents de la V^{ème} république. Dans ce corpus, l'auteur compare chaque sous-corpus (constitué des textes d'un président) au corpus global, avec une attention particulière au sous-corpus Chirac.
Le point de départ de notre analyse est la constatation que fait Mayaffre à la page 140, à propos de la forme *jeunes*. Cette forme est statistiquement la troisième forme la plus discriminante du discours chiraquien lorsqu'on le contraste sur les discours des autres présidents de la V^{ème} république. Il titre « plaire : les jeunes ». Selon lui, l'utilisation du mot *jeunes* a « trois fonctions politiques évidentes » qui sont :

« - d'abord le mot permet de dépasser les catégories sociales et partisanes habituelles. Les « jeunes » sans distinction de classes et de parti se trouvent considérés en bloc. […]
- ensuite, loin de la gratuité linguistique, la sur-présence des « jeunes » est le signe fort de la logomachie engagée par Chirac et Jospin. A partir de 1997, les jeunes sont devenus un enjeu de pouvoir politique patent entre le président et son premier ministre. […]
- ultime hypothèse : aux antipodes de l'innocence linguistique, la sur-présence de « jeunes » dans le discours peut relever des manipulations de la nouvelle communication politique. (discours émotionnel plutôt que rationnel pour attirer la sympathie, […] touche l'affect.)[1]»

C'est, en fait, la première remarque qui a motivé une étude plus approfondie de cette forme. La question qui se pose est de savoir si le signifiant *jeunes* lexicalise un thème comme pourrait le laisser penser la citation précédente (création d'une entité subsumant les différences entre les différents jeunes), s'il en lexicalise plusieurs en fonction des contextes (p.142 « "l'intégration des jeunes" apparaît comme un moyen lexical détourné pour traiter des catégories issues de l'immigration), s'il ne lexicalise qu'une partie d'un thème, ou s'il n'en lexicalise aucun.
Remarque : Rappelons à ce propos les mises en garde de Rastier[2] concernant les difficultés inhérentes au passage de l'analyse lexicale à l'analyse thématique :

« à la différence des lexèmes, les thèmes ne sont pas des signes, ni, corrélativement des unités du français : ils dépendent en effet d'autres normes que la langue. Si le lexème et le thème diffèrent aussi bien par le niveau que par le palier d'analyse, le premier étant un signe, et relevant de la morphologie et de la microsémantique, le second une unité du contenu au palier mésosémantique, il est clair que tout lexème n'est pas un thème.
Une analyse thématique qui en resterait au palier lexical compterait potentiellement autant de thèmes que de mots de la langue. […] On objectera que les thèmes sont ordinairement dénommés par un lexème. Mais ce lexème est simplement une lexicalisation privilégiée du thème. Et l'on pourrait fort bien rencontrer des thèmes sans lexicalisation privilégiée.[3] »

Rastier donne plusieurs exemples de ce phénomène, notamment celui de la molécule de l'ennui dans Madame Bovary, molécule dont seules certaines parties sont lexicalisées[4]. Un autre exemple est celui du nombril chez Flaubert qui est selon lui :

« une lexicalisation partielle d'un thème dans la correspondance, d'un motif dans deux romans, et d'un topos isolé dans la correspondance. »

Cette problématique de la lexicalisation des formes sémantiques doit être appréhendée dans la perspective d'une théorie des genres. On peut donner cette citation de Rastier qui permet de faire le lien entre ces deux problématiques, et qui nous permettra de passer à la présentation de notre corpus :

« Résumons, un lexème peut ne lexicaliser aucun thème, par exemple, le mot thème ne correspond à aucun thème dans le corpus romanesque que nous avons étudié, mais il peut aussi en lexicaliser plusieurs. Enfin, son lien avec le palier thématique est relatif à un discours (littéraire, médical, etc.), un genre, et un corpus. »

Nous revenons donc dans ce travail sur l'analyse du corpus Chirac. Il faut noter cependant, qu'à la différence de Mayaffre, nous avons réorganisé le corpus Chirac, non plus seulement en fonction de l'ordre chronologique, mais d'abord en fonction des situations et du mode de communication. Encore loin d'une typologie des genres constitutifs de la pratique présidentielle, nous avons distingué dans un premier temps trois sous corpus en fonction du public (international, national ou régional), eux-mêmes divisés en deux en fonction du mode de communication (interactif ou non) : interviews ou discours. Cette répartition des discours entraîne des résultats sensiblement

[1] Mayaffre, D., 2004, pp.140-143.
[2] Rastier, F.,« La sémantique des thèmes ou le voyage sentimental » www.revue-texto.net
[3] Nous pouvons peut-être même aller plus loin en proposant qu'un thème spécifique n'a, par définition, pas de lexicalisation puisqu'il est abstrait de tout domaine. Lexicaliser un thème conduit à l'indexer sur un domaine.
[4] Rastier, F., *Arts et sciences du texte*, pp.200-201

différents de ceux de Mayaffre. Ce constat était prévisible, on se souvient en effet que Brunet a montré la dominance des contraintes liées au genre sur celles liées à l'auteur :

> *« Car de toutes les forces qui s'exercent sur un texte, le genre semble la plus pesante et la plus pressante. Nous gardons le souvenir décevant d'une expérimentation, réalisée avec Charles Muller, qui avait consisté à étudier les 60 mots français les plus fréquents dans une dizaine de textes (romanesques, dramatiques ou poétiques) de Hugo, Lamartine et Musset. Abusé par les méthodes statistiques, l'ordinateur avait reconnu des différences et invitait naïvement à conclure qu'il y avait trois auteurs différents: un romancier, un dramaturge et un poète.*[1] *»*

Notre méthode consistera dans un premier temps à observer le réseau de collocations de la forme jeunes, à l'aide de la fonction «contexte » du logiciel Hyperbase. On appliquera ensuite la fonction thème aux résultats obtenus, afin d'obtenir la liste de termes les plus fréquemment associés à cette forme. Les collocations relevées devront faire l'objet d'une analyse afin de préciser les relations qu'elles entretiennent avec le mot pôle et de déterminer leur degré de corrélation. Les résultats de cette phase nous permettront d'avoir une première idée sur le ou les thèmes que lexicalise en tout ou partie la forme *jeunes*. Nous pourrons alors étendre les requêtes aux corrélats les plus intéressants.
Cette étude a donc deux objectifs, le premier concerne l'importance d'une répartition problématisée du corpus en sous-corpus. Le second consiste, à travers l'étude détaillée des contenus lexicalisés par une forme, à évaluer les concepts nécessaires à l'analyse et ceux qui peuvent disparaître dans une dégradation raisonnée de la théorie.

2. Analyses

2.1. Constitution des molécules sémiques, analyse micro et mésosémantique

2.1.1. Les cooccurrences, nébuleuse de signifiants

Dans un premier temps, nous observons les collocations de la forme *jeunes* dans le corpus global, puis dans chacun des sous corpus. Nous effectuons donc la requête contexte appliquée à la forme « jeunes », puis nous appliquons la fonction thème aux résultats obtenus. Les tableaux résultant de ce processus sont les suivants (voir tableaux page suivante)[2].
Les résultats proposés pour les sous-corpus, sans contredire la première hypothèse de Mayaffre vont cependant la relativiser. En effet, même si le terme « jeunes » permet d'envisager un groupe « en bloc » et de faire l'économie de la distinction entre des populations fort différentes et parfois clivées, il ne recouvre pas la même réalité en fonction des situations d'énonciation.

[1] Brunet, E., « Un texte sacré peut-il changer ? Variations sur l'Évangile. »
[2] Nous n'avons pas jugé utile ici de travailler sur une base lemmatisée qui aurait permis de ne retenir que les substantifs et aurait délaissé des expressions comme : « les jeunes diplômés ». L'inconvénient de notre choix est en revanche de retenir des expressions comme : « les jeunes démocraties », qui ne sont pas pertinentes et risquent de fausser les résultats. Une fois encore, l'analyse ou du moins le contrôle manuel reste nécessaire. Ce relevé des collocations n'est donc qu'une étape informelle, il convient ensuite d'aller vérifier à la main les sauts qualitatifs qui permettent d'établir les collocations.

Jeunes corpus total

151.78	1069	1069	JEUNES
24.16	517	137	FORMATION
22.41	64	40	QUALIFICATIO
21.69	42	31	FILLES
18.53	193	62	INSERTION
18.20	923	156	EMPLOI
17.46	3845	397	LEUR
15.96	14	13	DIPLÔMÉS
15.93	59	28	ALTERNANCE
15.90	98	37	APPRENTISSAG
14.71	24	16	GARÇONS
13.19	262	56	PROFESSIONNE
12.94	45	20	QUALIFIÉS
12.79	59	23	CONTRATS
12.21	423	71	EMPLOIS
12.16	59	22	RETRAITÉS
12.11	353	63	ÉCOLE
11.86	620	89	EUX
11.63	15	10	SORTENT
11.54	157	37	JEUNESSE
10.99	69	22	SCOLAIRE
10.79	60	20	FORMATIONS
10.39	58	19	MOBILITÉ
10.37	165	35	MÉTIERS
10.36	43	16	INSTALLATION
10.20	2577	223	ILS
10.11	195	38	JEUNE
10.04	164	34	MÉTIER
9.72	28422	1616	LES
9.69	180	35	ENSEIGNEMENT
9.65	77	21	PROFESSIONNE
9.42	137	29	ÂGE
9.32	18	9	INSÉRER
9.31	26	11	BOULANGERS
9.19	357	52	CHÔMAGE
9.18	126	27	QUARTIERS
9.02	59	17	ÉDUCATIF
8.51	321	46	EXPÉRIENCE
8.36	26	10	FERMÉS
8.27	227	36	AIDER
8.25	48	14	UNIVERSITAIR
8.09	184	31	GÉNÉRATIONS
7.86	58	15	FORMER
7.84	34	11	CLASSE
7.81	87	19	ÉLÈVES
7.73	17348	990	QUI
7.66	41	12	AÎNÉS
7.51	55	14	ADULTES
7.49	4780	322	SONT
7.44	1592	133	NOTAMMENT
7.42	93	19	ÉPANOUISSEME
7.34	50	13	ACQUÉRIR
7.30	350	44	TROUVER
7.08	125	22	UNIVERSITÉ
7.03	117	21	ÉTUDIANTS
6.95	557	59	ENFANTS
6.85	23962	1296	DES
6.83	565	59	ENTREPRISE
6.82	7950	483	PLUS
6.76	56	13	ACTIFS
6.60	31	9	APPELÉS
6.59	241	32	CHANCES
6.52	4488	292	AUX
6.46	45	11	SCOLAIRES
6.39	1570	123	VIE
6.35	227	30	CHANCE
6.28	1359	109	BEAUCOUP
6.24	691	65	NOMBRE
6.23	266	33	ÉDUCATION
6.19	34	9	ISSUS

Jeunes TI[1]

62.59	64	64	JEUNES
13.94	5	4	APPRENTISSAG
12.96	22	8	MILLIERS
11.34	35	9	JEUNESSE
10.73	56	11	FORMATION
10.38	5	3	UNIVERSITAIR
10.38	5	3	ÉLITES
9.77	15	5	ÂGE
9.65	10	4	PROFESSIONNE
8.68	7	3	MOBILITÉ
8.68	7	3	DITES
8.07	8	3	RESPECTIVES
8.07	8	3	FORMER
8.02	14	4	CENTAINES
7.81	53	8	SAINT
7.70	4	2	RAPPELONS
7.70	4	2	ENFANT
7.44	16	4	ÉTUDIANTS
6.83	5	2	RÉCIPROQUES
6.83	5	2	INDIFFÉRENCE
6.83	5	2	DIEU
6.83	5	2	ARTISTIQUES
6.83	5	2	ADULTES
6.77	11	3	TRAVAILLEURS
6.73	40	6	PÈRE
6.36	107	10	JAMAIS
6.18	22	4	LANGUES
6.15	13	3	APPELÉ
5.91	703	31	AUX
5.90	14	3	OFFRIR
5.72	25	4	BIENVENUE
5.67	7	2	PORTÉS
5.67	7	2	MANIFESTATIO
5.67	7	2	JOURNÉES
5.67	7	2	ÉTABLISSEMEN
5.67	7	2	BUENOS
5.67	7	2	BRÉSILIENS
5.67	7	2	AIRES
5.45	16	3	ARRIÈRE
5.39	41	5	ENTIER
5.39	41	5	ÉDUCATION
5.26	17	3	UNIVERSITÉS
5.26	8	2	JEUNESSES
5.26	8	2	ENTHOUSIASME
5.00	31	4	CONNAISSANCE

Jeunes EI

46.12	23	23	JEUNES
17.16	5	4	OFFICIERS
11.04	3	2	RELIGION
11.04	3	2	INTELLECTUEL
11.04	3	2	HUMANISTE
11.04	3	2	EXPORTER
11.04	3	2	AUBE
10.02	8	3	OPPOSITION
9.51	4	2	MALADIES
8.46	5	2	PAYSANS
8.46	5	2	MULTIPLIER
8.46	5	2	ÉCOLES
8.46	5	2	ATTENTES
7.71	13	3	MESSAGE
7.68	6	2	SAGE
7.68	6	2	DESTIN
7.07	7	2	SIDA
7.07	7	2	PORTÉE
7.07	7	2	ALLÉ
7.07	7	2	AFGHANS
6.87	16	3	CULTURELLE
6.58	8	2	ÉTUDE
6.58	8	2	COMMENCE
6.58	8	2	CIVILES
5.82	10	2	CULTURES
5.73	22	3	CONTACTS
5.52	11	2	REMARQUER
5.25	12	2	MODERNISATIO
5.19	26	3	ALLEMANDS
5.07	27	3	MAIN

[1] Convenons d'abréger par T les tribunes (mode de communication non interactif) et par E les entretiens. I, N et L renvoient respectivement à International, national et local.

Jeunes TL

100.27 615 615 JEUNES
19.22 29 26 FILLES
18.10 42 30 QUALIFICATIO
17.63 293 88 FORMATION
14.29 1999 267 LEUR
12.67 470 92 EMPLOI
12.23 15 12 GARÇONS
10.81 121 35 INSERTION
10.73 11 9 DIPLÔMÉS
10.64 32 16 ALTERNANCE
9.45 11 8 SORTENT
9.30 305 56 EUX
9.19 9 7 PÂTISSIERS
8.99 37 15 MOBILITÉ
8.75 26 12 BOULANGERS
8.62 44 16 FORMATIONS
8.57 113 28 MÉTIER
8.54 27 12 CONTRATS
8.37 67 20 APPRENTISSAG
8.27 17 9 ENCOURAGE
8.25 176 36 PROFESSIONNE
8.17 14 8 FERMÉS
8.00 137 30 MÉTIERS
8.00 109 26 JEUNE
7.79 31 12 LAURÉATS
7.73 46 15 SCOLAIRE
7.53 162 32 CHÔMAGE
7.24 161 31 TROUVER
7.23 1151 125 ILS
7.19 30 11 ACQUÉRIR
7.17 239 40 EMPLOIS
7.00 22 9 QUALIFIÉS
6.89 37 12 ÉDUCATIF
6.83 188 33 ÉCOLE
6.55 12609 911 LES
6.51 12 6 COLLÈGES
6.26 7136 542 QUI
6.06 115 22 AIDER
6.04 556 66 NOTAMMENT
5.88 52 13 TROUVENT
5.78 29 9 INSTALLATION
5.71 114 21 CHEZ
5.50 31 9 ADULTES
5.46 173 27 SALUER
5.38 214 31 NOMBREUX
5.31 66 14 ÉPANOUISSEME
5.29 150 24 EXPÉRIENCE
5.24 33 9 FORMER
5.19 22 7 ISSUS
5.06 41 10 CARRIÈRE
5.00 130 21 PROBLÈME

Jeunes EL

19.94 63 63 JEUNES
6.51 20 13 EMPLOIS
5.50 2 3 CROISADE
5.49 34 16 EMPLOI
5.08 15 9 QUARTIERS

Jeunes EN

49.29 141 141 JEUNES
9.28 5 5 INSÉRER
8.32 78 21 EMPLOI
8.30 4 4 OPTIMISTES
7.66 7 5 ANPE
7.32 14 7 JEUNESSE
7.19 3 3 MINITEL
7.19 3 3 KILOMÈTRES
6.72 16 7 INSERTION
6.70 332 46 LEUR
6.60 9 5 ALTERNANCE
6.58 6 4 APPRENNENT
6.44 13 6 ACTIFS
6.00 7 4 VOLONTARIAT
6.00 7 4 APPRENTISSAG
5.53 8 4 ÉTABLISSEMEN
5.51 12 5 TRAITER
5.43 61 13 EUX
5.35 5 3 TRADITIONS
5.35 5 3 MINORITÉ
5.35 5 3 DÉLINQUANTS
5.35 5 3 COMMENÇANT
5.22 49 11 FORMATION
5.13 9 4 RETRAITÉS
5.13 9 4 CONTRATS
5.13 9 4 CITÉ

Comme nous le disions en introduction, la segmentation du corpus en fonction des situations d'énonciation et du mode d'interaction entraine des différences de résultats dans chacun des sous corpus. Les nébuleuses lexicales entourant le mot pôle nous orientent vers des réseaux différents. Quelques premières remarques peuvent être effectuées : le mot *emploi*, omniprésent dans les sous-corpus national et local est absent du sous-corpus international. On observe une certaine unité des sous-corpus National et Local face au corpus International.
On peut facilement imaginer que les jeunes dont parle le président dans ses discours et entretiens internationaux ne correspondent qu'à une certaine partie de la jeunesse. La jeunesse française sur laquelle le président communique à l'étranger étant mise en parallèle avec celle des autres pays (Brésiliens, Allemands, Japon) doit être une jeunesse brillante.
Outre les différences, on constate aussi des points communs. La forme *formation* est présente que ce soit au niveau international ou aux niveaux national et régional.

Une fois de plus, une grande prudence est de mise en ce qui concerne les signifiés de ces formes. Une relance des fonctions « contexte » et « thème » sur *formation* nous à permis de remarquer

deux emplois différents. La formation aux niveaux national et local est une formation concrète, appliquée et professionnalisante, au niveau international, elle est surtout universitaire. Se pose ici la question de l'unité du corpus Chirac. Une unité présupposée entrainerait la suspicion d'une duplicité de la part du président, alors qu'on peut considérer qu'il entre dans des pratiques différentes en fonction des situations de communication, et que ces pratiques ont leurs doxa propres. Il en va sans doute de même pour la forme *jeunes*. (cf. aussi note [8])

2.1.2. Les corrélats : réseaux de signifiés

Une fois ce premier traitement exploratoire réalisé, nous devons retourner vers le texte et essayer de définir quelles sont les isotopies ou paratopies[1] qui permettent d'établir les corrélations.

Rappel : « l'hypothèse qui fonde la transformation de la cooccurrence en corrélation est celle-ci : le contexte proche est structuré par des isotopies qui marquent l'appartenance à un même fond sémantique, ou des paratopies – qui marquent l'appartenance à la même forme sémantique[2]»

Une première remarque concerne le sémème de jeunes. Mayaffre obtient des résultats différents en termes de listes, mais qui vont dans le même sens que les nôtres :

> *« La consultation des contextes d'utilisation de « jeunes » montre que le terme apparaît presque exclusivement dans un environnement lexical favorable. »*

Nous nous rangeons aussi dans une certaine mesure à se dernière remarque :

> *« Il s'agit dans le corpus du Chirac d'un thème et d'un terme positifs d'un discours quasi subliminal. »*

Ce terme paraît en effet ne lexicaliser qu'une valorisation /méliorative/, et une thymie euphorique. Nous ne sommes pas pour autant sûr qu'il lexicalise un thème. Si on relance la requête « contexte » sur la forme *élites* dans le discours international, toutes les occurrences sauf une se rapportent aux jeunes.

Observons maintenant ce que peuvent être les sèmes communs de ces cooccurrents. On a vu que les évaluations thymiques et axiologiques sont présentes dans les sémèmes de « jeunes » et « élites », ils le sont aussi dans ceux de « apprentissage », « formation », « universitaire » et « mobilité ». Autre sème commun à la plupart de ces lexèmes, le sème aspectuel /inchoatif/ que l'on retrouve dans « jeunes », « apprentissage », « formation », « emploi », « insertion », « optimistes », et enfin le sème mélioratif dans « formation » « apprentissage », « insertion ».

On observe donc un faisceau d'isotopies spécifiques qui constituent une partie du fond du discours. (Les isotopies /euphorie/, /inchoatif/ et /mélioratif/.). Une fois relevées, ces isotopies demandent à être articulées, et on peut donc tenter d'établir une molécule sémique. Cependant, le problème qui se pose est celui des limites de cette molécule et de sa lexicalisation. Ici, comme nous venons de le voir, plusieurs termes peuvent prétendre au statut de lexicalisation synthétique de la molécule. Nous proposons donc une représentation structurée des parties de leurs sémèmes qui sont lexicalisables par les autres.

Commençons par le sémème de « jeunes ». Il est ici considéré comme un processus inchoatif, valorisé de façon positive. On aurait donc une première ébauche de la molécule de la forme suivante :

---

[1] Définitions : L'isotopie est l'effet de la récurrence syntagmatique d'un même sème dans différents sémèmes. L'identité des sèmes entraine l'équivalence des sémèmes. La paratopie est constituée par la récurrence syntagmatique de sèmes différents, mais appartenant à la même molécule sémique.

[2] Rastier, F.,1996,« La sémantique des thèmes ou le voyage sentimental » 2.3.b.

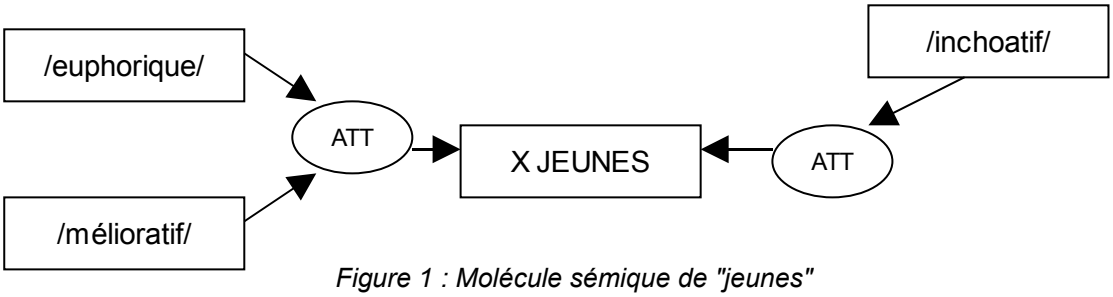


Figure 1 : Molécule sémique de "jeunes"

Une fois cette molécule établie, nous constatons que certains traits identifiés dans les autres termes ne sont pas présents. Nous devons donc les intégrer à cette représentation, quitte à la modifier. Le sème /mélioratif/ présent dans « formation » ou « insertion » doit par exemple être intégré. On peut faire évoluer cette molécule en remplaçant JEUNES par un sème /processus/ qui permet d'évacuer les sèmes génériques de JEUNES et d'abstraire la molécule d'un domaine particulier. Nous proposons la molécule suivante :

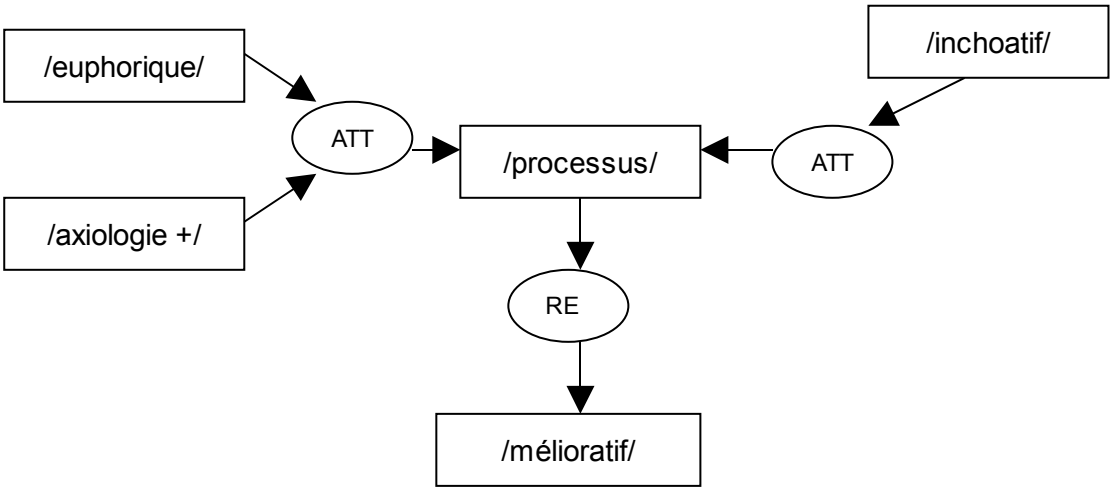


Figure 2 : molécule globale sans lexicalisation

Cette représentation nous permet de renchérir sur la remarque de Mayaffre concernant le « terme positif d'un discours quasiment subliminal ». Le procédé employé est assez semblable à celui de l'image subliminale : la forme est bien présente mais quasiment jamais totalement visible (forme bruitée). Nous revenons en revanche sur la première partie de son affirmation « un thème ». Il semble qu'avec les outils de la sémantique interprétative, on puisse affirmer que la forme jeunes ne lexicalise pas un thème tel que nous venons de le décrire, mais simplement un partie de thème, plutôt axée sur la valorisation et l'aspect inchoatif. La forme élites quant-à elle ne lexicalise que la partie évaluative de la molécule, la forme insertion semble être un bon candidat pour une lexicalisation synthétique.
Cette molécule semble lexicaliser un topos[1] quasi constitutif du discours politique et qui est celui qu'on pourrait nommer : « les lendemains qui chantent », présent dans presque tous les slogans de campagne (un nouvel élan…). Nous pouvons observer maintenant la lexicalisation de cette molécule dans les différents domaines du texte politique[2]. Toujours en partant du postulat que le

[1] Définitions : (glossaire de *Arts et Sciences du Texte*)
- thème (spécifique) : molécule sémique relevant du palier mésosémantique
- topos :
interne : enchaînement récurrent d'au moins deux molécules sémiques ou thèmes
externe : axiome normatif sous-tendant une afférence socialisée
- motif : structure textuelle complexe de rang macrosémantique. […] Ainsi le motif est un syntagme narratif stéréotypé, partiellement instancié par les topos.

[2] Remarquons au passage l'utilité que pourrait avoir une base dans laquelle les sèmes seraient déjà associés aux formes, il suffirait alors de relancer une recherche sur la collocation de ces sèmes pour voir comment la molécule prend chair dans les différents domaines. Faute d'une telle base, nous pouvons toujours relancer les requêtes sur les termes polysémiques de notre liste, ou sur certains synonymes susceptibles d'être présents dans le corpus.

contexte proche est structuré par les relations sémantiques, on peut étendre la recherche avec d'autres formes qui lexicalisent des parties de cette molécule. Cette étude permettrait en outre d'observer comment une même forme sémantique prend corps dans différents domaines et les différences en fonctions des sous-corpus. En effet, même si nous avons défini une molécule suffisamment générale (donc spécifique) qui peut rendre compte de ce que lexicalise la forme jeunes dans le corpus, il convient maintenant de voir quels sont les actants qui l'investissent dans les sous-corpus et comment elle s'articule avec les autres formes.

2.2. Analyse macrosémantique et textualisation des formes sémantiques

La macrosémantique relevant de l'ordre du texte et de l'intertexte, elle est le palier auquel on met en relation les différentes formes sémantiques. Ce palier est donc le lieu de l'articulation dialectique des formes sémantiques, celui de la description des motifs. Nous revenons ici vers les textes des différents sous-corpus pour commencer à envisager les différents investissements que peut prendre une même lexicalisation d'une même forme, dans des contextes différents. Le cas de notre étude paraît très illustratif. La figure 1 présentait la molécule sémique correspondant à la forme *jeunes*. Cette molécule est lexicalisée dans tous les sous-corpus par la forme *jeunes*, et pourtant elle renvoie à des réalités différentes car elle est contextualisée de façon différente et parce qu'elle s'inscrit dans des schémas narratifs différents.
Notons simplement que dans les sous-corpus local et national, l'état s'engage à assurer une formation professionnalisante qui garantira l'insertion des jeunes (en difficulté ou au chômage) :

> *« […]ouvrir le plus possible l'école et l'université sur l'entreprise. Il faut encourager, familiariser les garçons et les filles qui se forment avec le monde du travail. Ne pas se trouver ensuite avec une réponse faite à un jeune qui se présente et à qui l'on répond : " Mais quelle est votre expérience ? - Vous n'en avez pas, alors allez en acquérir et ensuite on vous donnera un emploi [1]". »*

alors qu'au niveau international, les jeunes sont déjà formés dans une grande mesure, l'apprentissage ne concerne pas ici l'alternance en entreprise, mais celui des langues pratiquées dans l'union[2]. Les jeunes du sous-corpus international correspondent à une population différente. Il s'agit là de l'avènement des élites formées dans les grandes capitales européennes selon les vieilles traditions du moyen-âge et de la renaissance.

> *« Il faut que nos jeunes, quel que soit leur parcours, deviennent européens en acquérant leur formation dans plusieurs villes universitaires de l'Union, renouant ainsi avec une tradition qui remonte au Moyen-âge.[3] »*

3. Conclusion

3.1. Remarques sur l'unité du corpus et l'importance d'une répartition problématisée des textes dans un corpus

On mesure dans cette étude l'importance de la répartition / organisation des textes dans le corpus. Les corpus, comme ensembles de textes, ne permettent pas toujours toutes les observations. Il semble aller de soi que l'étude des caractéristiques d'un discours ne peut faire l'économie des outils que propose la sémantique interprétative, notamment en ce qui concerne l'écologie des textes. Rappelons que d'après RASTIER :

> *« les textes sont configurés par les situations concrètes auxquelles ils participent ; en outre, par la médiation des genres et les discours, ils s'articulent aux pratiques sociales dont les situations d'énonciation et d'interprétation sont des occurrences.[4] »*

---

[1] Intervention de Monsieur Jacques CHIRAC Président de la République lors de la rencontre avec des jeunes des travaux publics à Villepinte VILLEPINTE - MARDI 22 AVRIL

[2] Une rapide observation révèle que tous les emplois de *apprentissage* concernent, au niveau international, l'apprentissage des langues, alors qu'ils concernent quasi exclusivement, au niveau local, l'apprentissage pratique d'un métier, l'acquisition d'un savoir-faire, d'une expérience dans le monde du travail.

[3] Discours de Monsieur Jacques CHIRAC président de la République à l'occasion de la réception des Ambassadeurs (Palais de l'Elysée) Palais de l'Elysée - Mercredi 26 août 1998

[4] Rastier, F. « Eléments de théorie des genres » (Texte diffusé sur la liste fermée Sémantique des textes, 2001) www.revue-texto.net

Mayaffre, dans son ouvrage, s'applique à étudier le discours présidentiel (comme l'indique le sous titre : *Jacques Chirac (1995-2003) et le discours présidentiel sous la V*^{ème} *république*.). Il contraste les textes de chacun des présidents sur ceux de tous les autres et pense ainsi pouvoir caractériser le discours de chaque auteur. Nous constatons cependant que le corpus Chirac ne présente pas une unité thématique, en ce qui concerne les référents des lexicalisations de formes sémantiques. Là où une interprétation sociologique ou politique pourrait tirer des conclusions sur les stratégies de communication d'un auteur, une analyse sémantique doit observer plus précisément les critères écologiques qui lient le texte à son entour, son genre, son discours, la pratique à laquelle il correspond, le champ dans lequel elle s'insère. Il semble donc que l'auteur soit un facteur faible de caractérisation des textes sans doute nécessaire, mais en rien suffisant, le style ou l'unité thématique ne devant être envisagés qu'au sein des genres. Une typologie des structures de corpus adaptée à une typologie des analyses reste sans doute un des chantiers de la sémantique voire de la linguistique de corpus.

3.2. Remarques sur l'analyse thématique et sur la lexicalisation des thèmes

Après avoir évoqué quelques recommandations concernant l'importance des conditions d'écologie fondamentales au sein de l'herméneutique, nous évoquons maintenant quelques remarques sur la construction des formes sémantiques. Nous reprenons ici les recommandations de Missire concernant les conditions herméneutiques nécessaires à la perception d'un thème latent. Cette perception est selon lui conditionnée par : « une lexicalisation synthétique dans l'entour proche où l'on identifie une lexicalisation analytique[1] ».
Dans nos textes, nous avons constaté qu'il y avait plusieurs lexicalisations synthétiques de la forme évoquée. Ce qui permet de repérer facilement le thème. Il n'en reste pas moins que le poids statistique d'une unité (en valeur absolue, ou relativement à d'autres corpus) ne garantit pas le statut de lexicalisation synthétique d'un thème. Comme nous avons pu le voir, la forme *jeunes* ne lexicalise qu'une partie de la forme.

3.3. Evocation du concept para synonymie thématique

Revenons une dernière fois sur les propositions de Régis Missire dans son chapitre consacré à la morphosémantique textuelle. Il fait, à propos des liens des molécules sémiques, le constat suivant :

> *« Si, en tant que facteurs structurants des molécules sémiques, les liens différencient les éléments qu'ils connectent, leur disparition doit pouvoir se comprendre comme une indifférenciation des valeurs instanciant ces nœuds. Cette indifférenciation est nécessaire pour permettre à chacune des parties de* valoir pour *la forme intégrale. »*

L'application de cette remarque conduit au concept présenté en titre. Les différentes lexicalisations analytiques se trouvent présenter des parties de formes évoquant la forme complète. Même si leurs sémèmes sont très différents, ils convoquent, au niveau thématique, la même molécule. Ces remarques restent encore trop générales, et de plus amples analyses sont nécessaires. Donnons à titre illustratif les deux exemples suivants :

- dans le sous-corpus National+Local, les termes *emploi*, *insertion*, *jeunes*, *formation*, *diplômés*, *sortent*… traduisent une même réalité ou un même topos qui est fondamental (pour ne pas dire fondateur) dans le discours politique et qui est celui de l'amélioration, de l'espoir (lexicalisable par exemple par : ça ira mieux demain, ou les lendemains qui chantent).
- dans le discours international, ce même topos est repris, la seule différence se situe au niveau du point de départ de l'amélioration. Là où le discours local propose l'amélioration d'une situation préoccupante, le discours international part d'une situation déjà excellente. Ici, les termes para synonymes sont les suivants : jeunes, apprentissage, formation, universitaire, élites, professionnel, mobilité, étudiants travailleurs…

On touche ici à une idéologie sous-jacente au texte. On a pu observer ici que cette idée, même si elle était constante dans les différents sous-corpus, ne se traduisait pas au moyen des mêmes actants. Nous voudrions pour finir, nuancer la première remarque de Mayaffre concernant

[1] Missire, R., 2005, *Sémantique des textes et modèle morphosémantique de l'interprétation*,Thèse de doctorat, Université de Toulouse 2, p.180.

l'utilisation de la forme *jeunes* (permet de considérer une classe en bloc et d'éviter d'aborder certains clivages). La diversité des réalités recouvertes par cette dénomination générale dépend peut-être d'une volonté de dissimulation de certains clivages, d'un œcuménisme facile, mais il dépend aussi sans doute des genres dans lesquels cette dénomination prend place.
Face à l'avènement d'une herméneutique numérique dans laquelle le texte interroge l'interprète, nous voudrions simplement rappeler que l'herméneutique est avant tout une question d'écologie des textes et que la tâche de l'interprète commence bien avant celle de la machine ne serait-ce que par le choix des textes à analyser. Rechercher l'unité d'un discours en ne différenciant pas les différents genres qu'il investit paraît illusoire. L'unité du discours ne pouvant s'appréhender qu'après avoir étudié les différents genres et fait la part de ce qui leur était propre.
Notre travail propose ici quelques recommandations méthodologiques préalables à l'étude des textes, il reste à une herméneutique critique (véritable épistémologie des sciences sociales) la tâche d'en faire le relevé complet et hiérarchisé.

BIBLIOGRAPHIE

BEAUDOUIN, V. 2000. Statistique textuelle : une approche empirique du sens à base d'analyse distributionnelle, www.revue-texto.net
BOMMIER-PINCEMIN, B. 2002. Sémantique interprétative et analyse de textes : que deviennent les sèmes ?, *Sémiotiques*,17, pp. 71- 120.
BRUNET, E. Un texte sacré peut-il changer ? Variations sur l'Evangile., http://magyar-irodalom.elte.hu/colloquia/000601/brunet2.htm
MAYAFFRE, D.2004. Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Vème république, Paris, Honoré Champion.
MISSIRE, R. 2005. *Sémantique des textes et modèle morphosémantique de l'interprétation*, Thèse de doctorat, Université de Toulouse 2, en ligne sur : www.revue-texto.net
RASTIER, F. 1987. *Sémantique interprétative*, Paris, PUF (édition de 1996).
RASTIER F.,1996, "La sémantique des thèmes ou le voyage sentimental", revue Texto ! www.revue-texto.net
RASTIER, F. 2001. *Arts et sciences du texte*, Paris, PUF.
RASTIER, F. 2003. Formes sémantiques et textualité, in *Unité(s) du texte*, Cahiers du Crisco 12, pp. 99-114.
RASTIER, F. 2005. Mésosémantique et syntaxe, revue Texto ! www.revue-texto.net

L'OEUVRE POÉTIQUE DE TUDOR ARGHEZI. LA DIVERSITÉ DU LEXIQUE ET LE PROBLÈME DU STYLE

Simona CONSTANTINOVICI
Université de l'Ouest, Timişoara, Roumanie

Depuis bien des années on a essayé d'établir les traits caractéristiques du vocabulaire de l'un des plus grands écrivains roumains : Tudor Arghezi. On a établi jusqu'à présent la richesse et la diversité du lexique de cet auteur en utilisant une banque de mots très importante, conçue à partir de ses oeuvres (poésies, romans et pamphlets). Par la thématique qu'il répand tout au long de son oeuvre, par ses conceptions littéraires, religieuses, politiques etc., il privilégiera inévitablement, comme tout grand écrivain, d'ailleurs, certains mots aux dépens des autres. Cet article se propose d'identifier la complexité du vocabulaire et de démontrer qu'il est finalement un indicateur sensible et fidèle du style. Il constitue en grande partie le style d'un auteur. Par l'intermédiaire du lexique, on pourra aussi établir l'influence exercée sur cette oeuvre par d'autres auteurs, roumains ou étrangers (Mihai Eminescu et Charles Baudelaire, par exemple).
On a décélé deux sortes de registres lexicaux inclus dans le lexique dominant d'Arghezi. Autrement dit, celui-ci comporte, en ce qui concerne l'oeuvre de cet auteur, deux facettes : les mots au sémantisme négatif, d'une part, et les mots au sémantisme positif, d'autre part. Bien qu'ils soient moins significatifs du point de vue numérique, les mots vulgaires, argotiques ont une grande capacité à générer une vraie puissance textuelle et des proximités lexicales des plus inattendues.
La stylométrie, telle qu'on l'envisage dans cette étude, montre que le lexique des poèmes de Tudor Arghezi se distingue sensiblement de celui des romans ou des pamphlets.

La diversité du lexique. Quelques précisions.
Cet article est un modeste hommage à l'un des plus grands créateurs roumains du XXème siècle : Tudor Arghezi.
Les linguistes roumains affirment, à peu près tous, que Tudor Arghezi est l'initiateur de la reviviscence du langage poétique roumain par l'assimilation de l'élément lexical multiple, d'une part vulgaire, unique dans ses résonances contextuelles, d'autre part doux, éminemment poétique. Finalement, tout grand auteur, celui qui dans l'histoire littéraire inscrit son nom dès ses premières lignes, fait inconsciemment une révolution dans la littérature.
Dans le dictionnaire que l'on a conçu, unique à travers la bibliographie de spécialité, on a voulu montrer comment un auteur pourrait se faire mieux connaître à l'intérieur de son pays et peut-être aussi à l'étranger. D'abord, par les ouvrages lexicographiques et ensuite, si c'est possible, par les traductions proprement dites qui ont sans doute besoin d'un dictionnaire des termes littéraires. Car le vocabulaire, sa diversité, ses empreintes traduisent finalement un style, ici celui de l'un des plus grands auteurs européens à notre avis.
La structuration du lexique dans *Les Fleurs de Moisissure*, l'oeuvre représentative de cet auteur, qui peut être considérée comme le correspondant des *Fleurs du Mal*, se réalise à différents paliers. En fait, ce type de vocabulaire est le résultat de la synthèse de tous les mots (entrées lexicales) existant dans une langue à un moment donné. Il possède un aspect encyclopédique, car il rassemble plusieurs registres : archaïque, populaire, religieux, scientifique, régional etc. Le lexique est plein de nuances, point de départ de suggestions multiples. Il suffit de savoir l'activer, l'observer attentivement.
L'objectif de notre recherche, qui sera limitée à une part minime mais représentative de l'ensemble de l'oeuvre d'Arghezi, est de dégager les relations, les affinités stylistiques qui sont de trois sortes : 1. des triades synonymiques ; 2. un mot simple et les mots qui en sont issus, soit par dérivation, soit par composition (il y a dans cette catégorie des mots nouveaux, on pourrait les nommer inventés par Arghezi) ; 3. des sémantismes temporels et de l'espace.
Nous avons donc essayé de mesurer l'efficacité, les apports et les limites de la méthode lexicographique en matière non seulement d'étymologie, de formation des mots et de relations sémantiques, telles que la synonymie, l'homonymie, la polysémie, mais surtout de diversité d'un lexique poétique dans toutes ses articulations. Il y a des mots, des entrées lexicales qui reviennent avec une fréquence remarquable. Cela traduit le goût de l'auteur pour un certain registre lexical ou stylistique. Par exemple, il n'y a pas de mots proprement religieux, sauf ceux des psaumes, qui

sont des créations religieuses par excellence. *Les Psaumes* activent un registre lexical religieux. Pour cette spécificité, on s'arrêtera sur un vocabulaire de ce type[1], enregistrant :
1. les *personnages bibliques et mythologiques* (on n'inclut pas à ce niveau-là les noms propres) : *ange* (34) provient du latin *angelus*, un mot commun, susceptible d'être utilisé plusieurs fois dans la poésie moderne et qui devient dominant par fréquence / potence stylistique ; dér. *angélique* (une seule apparition dans le texte) « Mon ongle *angélique* s'est épuisé [...] / Je l'ai laissé croître » ; *apôtre* (2) ; *archange(s)* (2) ; *chéruvim* (4) ; *mage(s)* (1).
Ces entrées lexicales, que l'on pourrait considérer comme des *mots internationaux*, étant donné leur origine connue, sont issues de l'Antiquité latine.
2. la *divinité*. Le nom de la divinité est écrit toujours en majuscule dans le texte poétique d'Arghezi. Ceci est sans doute naturel pour un créateur qui s'est trouvé sa vie entière en dialogue avec la divinité. *Seigneur* « Doamne » apparaît 102 fois dans sa lyrique, et Dieu « Dumnezeu » 84 fois. La plupart des occurences sont au vocatif.
3. les *éléments de culte* : la *bible* (paradoxalement, il n'y a qu'une seule apparition dans la poésie), *bénir* (2), la *bénédiction* (7).
4. les objets de culte (vêtements, bouquins, etc.) : la *soutane* (1), mot d'origine turque qui désigne un habit spécifique aux prêtres orthodoxes.
5. les *composants de l'architecture écclésiastique* : l'*église* (26), la *chaire* (1). Il y a peu d'éléments lexicaux qui nomment ce registre. Sauf le vocabulaire religieux perçu comme général, très connu : *ciel, croix, Dieu, éternité*, il y a un lexique non-religieux, commun, qui manifeste la tendance de conversion.

À un autre niveau d'intérêt se situe le néologisme[2]. Apparemment indifférent à la place occupée dans une lignée synonymique (lexicale ou seulement contextuelle, métaphorique), le néologisme apporte une nouveauté expressive incontestable. Au-delà de l'innovation lexicale, pour mieux évaluer le contenu, il faut prendre en considération que nous n'avons affaire qu'à la première partie d'une longue entreprise. Il ne faut pas ignorer le lexique d'une autre nature repérable dans le texte poétique, argotique, archaïque etc. L'étude de ces mots et des contextes poétiques qu'ils engendrent peut devenir un lieu d'intérêt pour les stylisticiens et les sémanticiens.
L'étude a été réalisée sur un échantillon d'environ 10 000 mots issus de la poésie. Par la suite, on se propose de prendre également en considération les volumes de prose et d'écriture journalistique.
Le lexique poétique permet une analyse détaillée de la polysémie, dans la mesure où les mots réalisent, dans ce genre de texte, des transferts[3] qui élargissent la sphère sémantique et tracent de véritables constellations de sens contextuels, momentanés. La lexicologie considère comme modalités sémantiques essentielles : la métaphore, la métonymie, la personnification et la synecdoque. La polysémie devient ainsi, comme la synonymie, d'ailleurs, l'une des caractéristiques de la grande création.
Parmi les figures de style, les épithètes et les comparaisons sont les plus utilisées, elles configurent la spécificité de ce texte poétique et montrent en fait le sémantisme temporel et celui de l'espace. Seul un utilisateur raffiné des mots, seul celui qui possède l'art d'utiliser les mots, est susceptible d'engendrer la Poésie. C'est indubitablement le cas d'Arghezi.

Les psaumes d'Arghezi. Le problème du style
Les deux coordonnées de l'existence humaine, celles qui nous placent dans le monde, le temps et l'espace, connaissent au long de l'histoire littéraire et philosophique, des interprétations différentes, en fonction de la priorité que nous accordons à l'une ou à l'autre. Le sémantisme des périodes temporelles implique, de la part de celui qui *est* dans le monde, une autoperception et une autoconnaissance de sa propre condition. Pour Tudor Arghezi, le temps existe ; il est plus que l'espace, il est tout et en lui « se verse » Dieu. La temporalité domine la substance textuelle. Quoi que nous puissions dire, le sémantisme de l'espace, le monde d'Arghezi, le poète, n'est pas très différent de celui de Pascal, le philosophe ; seulement l'infinité débordante qui enivre le poète,

[1] Simona Constantinovici, *Dicționar de termeni arghezieni*, vol. I, Literele A-F, Timişoara, Editura Universității de Vest, 2004.
[2] G. I. Tohăneanu, *Dicționar de imagini pierdute*, Timişoara, Editura Amarcord, 1995.
[3] François Rastier, *Sens et textualité*, Paris, Hachette, 1989.

effraie le philosophe. Mais, tous les deux cherchent Dieu, sans se proposer du tout de nous délivrer de la peur devant les divinités. Etant donné que la poésie d'Arghezi est centrée sur la quête de la divinité, en quelque sorte de la *prière* et du *chant* religieux modernes, figures lexicales éminemment répétitives, elle peut être considérée comme la poésie de la répétition par excellence. Nous pouvons intégrer aussi le sémantisme des périodes temporelles et celui de l'espace dans le même ensemble stylistique de la répétition. Comme nous avons tenté une brève analogie entre la pensée poétique d'Arghezi et *les Pensées* de Blaise Pascal, il serait intéressant de montrer comment l'espace et le temps deviennent sources profondes et, en quelque sorte, inhérentes, du paradoxe[1]. Nous tenterons de le démontrer à l'aide des exemples choisis, en tenant compte du fait que l'art, en effet, tel qu'Arghezi le conçoit, est la négation de tout système philosophique puisqu'il le remplace[2]. L'immensité de l'espace s'oriente vers l'atemporalité. Nous sommes plongés dans *illo tempore*, *ab origine* dès les premières questions posées dans les psaumes à une divinité qui refuse constamment de montrer son visage. Plus loin, dans les mêmes poèmes, Arghezi atteint une philosophie de l'espace. Pour notre poète, l'espace humain est ce qui se distribue autour d'un centre. Et ce centre ne peut être que Tudor Arghezi lui-même. Dans cet univers, chaque objet et chaque être peut donc devenir le centre du cercle. Il y a dans son espace, un état visqueux de la pensée et du corps[3]. Selon notre poète, les vers doivent subir les lois de la mesure, fait qui montre combien Arghezi a été attiré par le souffle classique. Que signifie *l'abîme* pour Arghezi ? Quelque chose qui est concomitant avec notre existence. Il est le double, l'ombre de l'existence. À travers une expérience rationnelle, le passage, dans la pensée, de l'existence habituelle au néant (abîme) suppose une rupture entre deux niveaux de la connaissance : *être* / *paraître*[4]. En vertu de sa croyance et de son inspiration, les abîmes ne peuvent pas être définis en termes habituels, mais ils acquièrent la force de soutenir et d'engloutir l'esprit du poème. « Mesure, comprend l'espace » dit Arghezi, mais sa prière-exclamation, sa prière-négation, n'est pas conforme à l'aspect extérieur, matériel de son psaume. Nous avons devant nos yeux le psaume des temps modernes qui cherche en vain à s'approprier la mesure et la forme parfaites. Après qu'il se penche (s'être repenti), en suppliant la mesure suprême, Arghezi revient en disant que le poème est un signe qui subit les lois de la répétition. Et nous voyons, par la suite, comment Arghezi demeure fidèle aux nouvelles théories du signe poétique et nous comprenons, ainsi, mieux peut-être, pourquoi nous assoçions souvent la poésie de Rimbaud ou de Mallarmé à l'âme de la pensée d'Arghezi. Il y a, à ce niveau-là, certainement, une influence inavouable, sous-jacente. Le poème, pour Arghezi, est « un acte imaginaire, créant / le temps nécessaire à sa résolution. ». Ainsi, le temps est une entité ambiguë, soit dilatée, soit raccourcie, selon les dimensions de l'espace qu'il occupe. Arghezi est essentiellement le poète du simultané. L'ambiguïté sémantique peut intervenir tout de suite, à la force d'un mot intercalé au niveau le plus habile de la phrase, là où nous ne nous y attendions pas, car lui n'est nulle part démonstratif, au contraire, il nie partout la certitude scientifique. Le mot *seuls*, par exemple, entre dans le syntagme « l'un avec l'autre », une paradoxale association qui mène vers les couples consacrés : *Dieu – poète* ; *muse – poète* ; *mer – poète* etc. Pour accéder au ciel pur, il faut dépasser le niveau du bas, où tout est peint aux couleurs de la terre. La poésie de Tudor Arghezi est tantôt primitive, tantôt raffinée, enterrée dans la sensualité, dans la réalité des choses, lourde et inerte ou, au contraire, angélisée, éthérique. Le poète a l'ambition sacrée d'engloutir toutes les sensations que lui offre le monde ouvert. Il regarde et il pèse et, au bout de la contemplation et de la palpation, il crée le poème en tant que tel. *Le soleil, l'eau, la terre, le feu* sont les signes de ce monde anodin où nous sommes et, en même temps, le symbole de l'au-delà. Ils confirment, indirectement, l'adhérence à un espace. *L'eau*, élément incertain, associé dans les psaumes à l'esprit, c'est un autre signe poétique qui construit le texte et qui fait partie de l'espace imaginaire dans lequel nous vivons, comme êtres humains. Les psaumes contiennent des vers dans lesquels le thème du labyrinthe apparaît comme une protubérance poétique. Ce qui frappe à

[1] Pour M. Heidegger, par exemple, le problème de la parenté entre la poésie et la pensée à travers la médiation du langage est le problème-source de son système philosophique.

[2] Arghezi oppose la foi à l'art. C'est une opposition de circonstance car les deux ont des affinités.

[3] Georges Poulet compare cet état de viscosité et de pesanteur de la pensée et du corps avec l'état caractéristique aux romans sartriens. (*Oeuf, Semence, Bouche ouverte, Zéro*, p. 448).

[4] Arghezi a eu le sentiment de cette rupture développée et exprimée, pareil au sentiment de l'incompréhensible grandeur de l'homme. Pour lui, à un certain moment, se manifeste, à partir de cette rupture, la captivité volontaire de Dieu.

la lecture, étant donné que Tudor Arghezi est un chrétien atypique, donc qu'il croit en *un seul Dieu*, c'est la forme *mes Dieux*, qui accentue le paradoxe de sa haute écriture, sinon de sa pensée. Mécontent de l'image que lui offrait une seule divinité, l'auteur invoque la pluralité. Pour s'exprimer complètement, le poète a eu besoin de plusieurs voix protectrices, voix qui, d'ailleurs, existaient en lui, sous la forme la plus commune du *dialogue intérieur*. Nous savons maintenant que l'expression littéraire la plus évidente pour exprimer ses paradoxes (la plupart ontologiques) a été, pour Arghezi, le dialogue, sous la forme *duo* ou *duel*. Un peu plus loin, un autre signal désespéré, témoigne d'une certaine inconstance de la croyance chrétienne ou, plutôt, d'une certaine tendance à s'échapper. Il préfère affirmer avant de développer, d'où l'écartement de la décantation des choses. Il se pose tout le temps des questions, mais il préfère ne pas en donner la réponse. *Bénédiction* est un mot qui pèse beaucoup à l'intérieur du vers d'Arghezi, d'une part par sa longueur (*le signifiant*) remarquable, qui se répète six fois, deux fois dans le voisinage d'un autre mot long comme *accroissement* et, d'autre part, par sa signification (*le signifié*). Et même le croyant peut devenir inquiet, il peut demander l'impossible, ce que nous ne devons pas exiger : « Combien de temps encore ? », mais ce à quoi nous pensons tout le temps, comme si nous étions prisonniers d'une quête imaginaire. L'incertitude (*combien ... encore*) se transforme en méconnaissance, augmentée par la fréquence avec laquelle l'auteur utilise les mots négatifs, tels que *moisissure*, *ténèbres*. Et nous revenons constamment à la réflexion de Nietzsche, même si nous ne le disons pas directement : « La foi chrétienne est essentiellement un sacrifice, sacrifice de toute liberté, de toute fierté, de toute confiance de l'esprit en soi-même ; elle est en même temps asservissement et dépréciation de soi-même, mutilation de soi-même. Il entre de la cruauté et du phénicisme religieux dans cette foi qui se propose à une conscience fatiguée, complexe et blasée ; elle implique que la soumission de l'esprit soit inexprimablement *douloureuse*, que tout le passé et les habitudes d'un tel esprit se rebellent contre le comble d'absurdité qui s'offre à lui sous le nom de "foi". »[1]. Le paradoxe s'élève même sur l'accent ironique des mots. Les psaumes d'Arghezi traînent les mots jusqu'au bout de la question primordiale. Impétueux, le style d'Arghezi illustre la grande diversité du signe poétique. Et le signe poétique s'accompagne souvent, chez Arghezi, de *blancs* (marqués ou seulement imaginaires), véritables points d'une réflexion profonde, très importants, sans doute, et qui constituent la première différence visible entre la poésie et la prose. Avant de proclamer le signe poétique, le poème sort du silence antérieur, oscillant entre *la présence* et *l'absence*. Les structures répétitives, les sémantismes négatifs et positifs jouent leur rôle dans ce jeu linguistique infini.

BIBLIOGRAPHIE

(Ouvrages critiques)
ALEXANDRESCU, S. 1966. *Simbol şi simbolizare*. *Observaţii asupra unor procedee poetice argheziene*, în *Studii de poetică şi stilistică*, Bucureşti, Editura pentru Literatură.
BALOTĂ, N. 1979. *Opera lui Tudor Arghezi*, Bucureşti, Editura Eminescu.
BARTHES, R. 1953. *Le degré zéro de l'écriture*, Paris, Editions du Seuil.
CHEVALIER, J. et GHEERBRANT, A. 1989. *Dictionnaire des symboles*, Paris, Robert Laffont Jupiter.
CLAUDEL, P. 1957. *Oeuvre poétique*. Paris: Editions Gallimard.
CONSTANTINOVICI, S. 2004. *Dicţionar de termeni arghezieni*, vol. I, Literele A-F, Timişoara: Editura Universităţii de Vest.
*** *Dicţionarul explicativ al limbii române*, ediţia a II-a. 1996. Bucureşti: Univers Enciclopedic.
*** *Dicţionarul limbii române* [publicat de Academia Română, sub redacţia lui Sextil Puşcariu]. Bucureşti. 1913-1949.
NIETZSCHE, F. 1989. *Par-delà bien et mal*, Paris, Gallimard.
POULET, G. 1987. *Du romantisme au début du XX-ème siècle*. Paris: PUF.
POULET, G. 1955. Oeuf, Semence, Bouche ouverte, Zéro, *Nouvelle Revue Française*, 33.
RASTIER, F. 1989. *Sens et textualité*, Paris, Hachette.
RASTIER, F. 1991. *Sémantique et recherches cognitives*, Paris, PUF.
TOHĂNEANU, G. I. 1995. *Dicţionar de imagini pierdute*, Timişoara, Editura Amarcord.

[1] Nietzsche, *Par-delà bien et mal*, p. 64.

UN INSTRUMENT DE LECTURE ANALYTIQUE : PRÉSENTATION DE *CORPUTEX*

Pierre SADOULET
CIEREC/Université de Saint-Étienne

SOMMAIRE

Résumé : *Quelle que soit la qualité du texte obtenu à la suite de la numérisation, ce texte n'en devient pas pour autant, comme magiquement, le révélateur de ce qu'il est. Pour le décrire et l'analyser, comme dans un travail sur livre, il faut pouvoir le lire et le relire non pas par fragment mais comme un tout – le global conditionne le local – et comme un tout dont chaque détail peut être très important. De plus un travail analytique doit repérer des récurrences ou des localités qu'il faut savoir identifier, analyser et classer à partir de la seule chaîne de caractères qui le signifie.*

Le travail de linguistique, de lexicologie et de sémiotique textuelle qui est le nôtre nous a conduit à privilégier l'activité herméneutique du philologue aux automatismes mécaniques de la machine et aux comptages porteurs d'une objectivité qui peut être illusoire. La machine doit offrir toute sa puissance à accélérer des tâches répétitives nécessaires (création de balises, lemmatisation, distinction des homonymes, indexation) ; mais elle ne peut remplacer le talent herméneutique de l'analyste qui doit, dans tous les cas, examiner le passage et décider de ce qu'il perçoit de son effet de sens, même si ses pesées toujours trop rapides font nécessairement l'objet de constantes révisions.

L'élaboration par l'Université de Irvine d'un corpus presque complet de la littérature grecque, le TLG, qui existe depuis plus de 20 ans, puis le remarquable travail de l'Atlif sur le corpus de Frantex ont constitué pour nous des moyens considérables. Mais les applications qui servent à les consulter nous ont conduit très rapidement à regretter une insuffisance qui nous gênait beaucoup : les fragments que ces applications extraient restent toujours trop courts et lorsqu'on retravaille les passages ainsi conservés, il faut toujours aller chercher le livre ou le texte pour retrouver le cotexte, ce qui conduit à différer la réponse au besoin créé par le moment de lecture particulier, voire souvent à renoncer à cette enquête nécessaire donc à risquer de laisser passer l'effet de sens dans toute sa complexité.

Corputex est une application écrite dans le progiciel de base de données 4D qui utilise l'interface et les fonctions de traitement de chaîne propres à ce logiciel d'une grande puissance pour proposer un instrument de travail qui permet de répondre à tous les besoins d'étude et d'analyse d'un texte long numérisé, tant au fil du texte qu'à l'intérieur de dossiers d'extraits. Un système de signets, de notes diverses permet d'annoter le corpus à tous les niveaux, y compris par l'extraction de citations. Une fois certaines occurrences marquées par l'analyste ou retrouvées par le logiciel (recherche lemmatisée, distinction d'homonymes avec une interface très rapide, recherche de concordances) les passages considérés peuvent être classés et commentés selon les besoins. Mais à chaque moment, il suffit d'un clic pour pouvoir retrouver tout le cotexte, qui est toujours disponible à l'affichage dans la base, si possible selon la même linéation et la même pagination que l'édition originale.

Ce logiciel par certains côtés peut apparaître comme une usine à gaz, de par la multiplicité des fonctions qu'il offre. Mais il nous a rendu de nombreux service d'abord dans des études sur le grec ancien mais aussi dans des travaux de linguistique et de sémiotique françaises qui demandaient de

repérer des passages représentatifs et des extraits permettant de mieux identifier au fil du discours certaines spécificités du fonctionnement sémantique et linguistique du corpus. Nous examinerons un cas particulier qui montrera l'apport de cet environnement logiciel.

Depuis 1980, année où j'ai passé mon doctorat de troisième cycle en morphosyntaxe du grec, j'ai occupé sinon perdu des années de travail avec l'idée que la possibilité de numériser un texte allait entraîner d'énormes progrès pour les sciences du langage, dans la mesure où l'informatique permettrait d'accélérer considérablement les procédures de mise en fiche. En effet, travaillant à une d'étude morphosyntaxique synchronique du grec ancien, j'avais passé des mois entiers à extraire du texte des passages pertinents, que je fichais selon une certaine méthode pour me permettre de repérer d'un simple coup d'œil les cas les plus intéressants. Je mettais ainsi en évidence des relations de commutation entre plusieurs constructions syntaxiques qui confirmaient empiriquement la description morphosyntaxique que je finissais par proposer.
De premiers essais informatiques m'ont donc conduit à une perte de temps considérable puisque je rêvais alors – erreur de jeunesse – d'un système informatique qui ferait, grâce à un travail de manipulation considérable, une sorte de préanalyse des énoncés offrant une visualisation directe des constructions, à l'image des fiches que j'avais confectionnées pour la thèse.
Comme j'étais alors professeur du secondaire et que j'avais, à côté de mon métier, une grosse activité militante, le temps utilisé pour toute ces recherches était celui des vacances et, faute des informations nécessaires, j'ai commis des erreurs fondamentales au niveau du choix des logiciels. Plus tard, sur les conseils de Richard Goulet, un collègue helléniste qui avait conçu un premier logiciel d'analyse lexicale et de lemmatisation (*Lexis*), écrit d'ailleurs avec un système Pascal, j'ai fini par adopter l'application 4D, un progiciel de gestion de bases de données liées, afin de profiter au mieux des facilités techniques de ce progiciel, notamment par la diversité des fonctions de traitement de chaîne et de recherche qu'il offre et par la possibilité qu'il donne, grâce à un traitement de texte interne, intitulé *Write*, de produire des documents mis en forme et lisibles par *Word*.
L'objectif avait évolué alors : il ne s'agissait plus d'avoir un analyseur plus ou moins automatique, qui risquait de mémoriser dès le départ un mode de formalisation syntaxique qui serait ensuite retrouvé quasi sûrement par la théorie. Je voulais une application qui ait le moins possible d'intelligence artificielle. Par contre, elle devait offrir une interface aussi rapide que possible pour formater et garder à disposition de l'utilisateur le corpus dans son entier.
C'est ce qui m'a conduit à écrire *Corputex*.
Corputex est d'abord conçu comme un logiciel de recherche de chaînes semblable à *Frantex*. Des moyens de recherche par chaîne, par lemme ou, tout simplement, par balisage direct au fil de la lecture permettent de retrouver les passages pertinents dans le cadre d'une étude donnée : ces « extraits » réunis par le logiciel dans des « dossiers » peuvent être ensuite analysés un à un, par la mise en place de classements et de commentaires à partir d'ensembles de propriétés hiérarchisées établies par le chercheur. Cet étiquetage par clé des extraits, toujours à revoir, peut se faire, au départ, très rapidement, grâce aux possibilités de l'interface. Il est ainsi fait appel à l'intuition immédiate de l'analyste qui, bien sûr, devra revoir ensuite plusieurs fois les extraits les plus caractéristiques, pour affiner les distinctions. Il finira enfin par sélectionner les plus intéressants qui lui serviront pour un travail approfondi avec, à ce moment-là, surtout, une évaluation attentive du *cotexte*. La tâche sera d'assurer définitivement les effets de sens des extraits dans leur contexte grâce au recours qu'il est possible de faire à l'ensemble du corpus et aux aides-mémoire qui y auront été aménagés.
Même si *Corputex* sait compter tout ce qu'il trouve, il ne fait jamais de statistiques compliquées. Car tout ce qui est fait sur *Corputex* est le résultat d'un jugement posé par son utilisateur. Chaque base de donnée, consacrée à un texte et à une étude particulière, est donc conçue strictement comme un instrument de travail individuel. Ce qu'il contient, c'est un corpus lu et annoté par un individu.
Il n'est pas question ici de faire une présentation ou une démonstration du logiciel mais d'en décrire les fonctions essentielles, tout en indiquant les bases théoriques qui ont conduit à son élaboration. Nous verrons donc d'abord son cahier des charges. Puis nous examinerons quelques fondements qui justifient les services qu'il peut rendre au professionnel des sciences du langage. Pour conclure, nous évoquerons trois expériences d'études faites avec l'aide de *Corputex*.

1. La conception du cahier des charges
1.1. Les insuffisances de l'existant

* Le TLG : le logiciel *Pandora*

Lorsque, après mon troisième cycle, dans les année 80, je m'orientais vers ce travail d'élaboration d'interface logiciel, il commençait à être diffusé, pour les hellénistes, un instrument de travail très précieux, le *TLG* (*Thésaurus linguae graecae*), un CD élaboré par l'Université d'Irvine[1], qui contenait l'ensemble des textes connus de la littérature du grec ancien. Ce corpus – critiquable du fait qu'il avait dû, pour des raisons de droits, reprendre des éditions anciennes des auteurs – n'en représentait pas moins un instrument d'investigation très puissant.

D'abord réservé au système d'ordinateur Ibycus, le corpus devint très vite accessible à tous les micro-ordinateurs *Apple*, grâce au logiciel *SNS Greek* de l'Université de Pise, et au système *Pandora*, développé par l'université de Chicago[2]. D'autres ont été développés depuis pour le système *Windows*[3].

Ces logiciels offraient la possibilité d'extraire des œuvres entières. Mais la recherche d'occurrences à partir d'un traitement de texte s'avérait difficile, en raison du codage particulier des jeux de caractère grecs accentués. Il fallait donc utiliser les logiciels eux-mêmes pour extraire les passages contenant telle ou telle chaîne de caractère. Un système de codage très complexe permet, dans ces systèmes, la recherche des séries de formes correspondant à une déclinaison ou à une conjugaison. Avec une bonne technique, il était donc possible de retrouver toutes les variantes morphologiques d'un lemme. La série d'occurrences ainsi retrouvées était exportée sous la forme d'une suite d'extraits du corpus contenant au plus quatre lignes avant et quatre lignes après.

Il était possible ensuite de traiter ces données soit dans un traitement de texte soit, après importation, dans des bases de données spécifiques[4].

* *Frantex*

Près de dix ans après, j'ai pu avoir accès à un autre système largement utilisé par tous les spécialistes de langue française : il s'agit de *Frantex,* la base de donnée de l'ATILF, disponible sur *Internet*[5].

Il faut dire que nous retrouvions le même défaut : même si nous augmentions au maximum le nombre de lignes copiées pour l'export, les extraits restaient parfois insuffisants pour permettre certaines identifications nécessaires à la bonne interprétation du passage. Or c'était justement ces données qui semblaient les plus décisives : c'est quand une occurrence résiste à l'interprétation immédiate qu'on peut penser avoir un cas qui permette d'améliorer le système descriptif utilisé. Que ce soit en morphosyntaxe ou en lexicologie, et même en sémiotique, ce sont les contre-exemples apparents qui peuvent faire avancer le modèle d'analyse[6].

1.2. De nouveaux objectifs pour un outil commode

Partant du constat de ces insuffisances très pratiques, j'ai donc essayé de les dépasser pour écrire l'application *Corputex*. Son cahier des charges, d'abord élaboré à partir de ces expériences, s'est complété peu à peu au fil des besoins.

1.2.1. Retrouver l'oiseau rare qui permettra de comprendre enfin

L'objectif de départ du projet est d'abord de faciliter un travail en linguistique. L'axiome premier de la démarche scientifique était qu'il fallait trouver un moyen de renouveler la description linguistique dans les nombreux cas où elle s'avère insuffisante en cherchant « l'oiseau rare », l'exemple

[1] réf. : [s.a.], 1991 – *Thesaurus grec de Irvine*, Californie. CD ROM.

[2] Il s'agit d'un pile Hyperbase pour le système Macintosh. Dernière version connue : 2.5.2. Voir l'aide au logiciel http://www.lib.uchicago.edu/e/ets/TLG.html .

[3] Si nous arrivons à résoudre certains problèmes de transcodage et d'identification de l'organisation du CD, il est techniquement possible d'avoir une version de *Corputex* qui gère directement les données du CD.

[4] Voir SADOULET Pierre, 1996

[5] Adresse actuelle : http://atilf.atilf.fr/frantext.htm. Nous rappelons que cet accès n'est disponible que sur abonnement.

[6] De plus, dans *Frantex*, avant la mise au point d'un corpus catégorisé permettant de différencier les homonymes, les recherches donnaient des résultats trop importants, ce qui obligeait ensuite à un fastidieux travail d'élimination.

inattendu qui serait l'occasion d'identifier une propriété particulière, le cas qui nous révèlerait, pour ainsi dire, qui nous ferait trouver le fonctionnement caché rendant compte de toutes les virtualités d'emplois de telle ou telle unité linguistique[1].

1.2.2. Extraire tout ce qui est pertinent par lecture directe ou par recherche de chaînes du signifiant

La recherche de ces « oiseaux rares » supposait d'abord des lectures et des sélections depuis le corpus entier. *Corputex* nous permettait de lire directement l'ensemble du texte. Les extractions pouvaient se faire à l'ancienne, en marquant scrupuleusement les passages intéressants du corpus par des balises qui permettaient à *Corputex* de les retrouver pour les copier ensuite dans un dossier.
Mais, souvent, il s'avérait qu'il était plus rapide de réunir les extraits à partir de recherches sur le signifiant, quitte à éliminer par un simple clic les occurrences non pertinentes. Il a fallu donc d'abord concevoir un système de recherche par chaîne qui puisse exiger une correspondance des codes ascii, pour tenir compte des caractères accentués et des majuscules. Puis j'ai pu mettre au point des algorithmes spécifiques, qui existent aussi dans *Frantex*, pour des recherches par mots ou même pour identifier des concordances et autres collocations.

1.2.3. Lemmatisation et distinction rapide des homonymes

Une indexation automatique permet de mémoriser toutes les occurrences des mots du corpus, y compris en traitant les mots composés avec tirets et en réunissant les formes ayant une césure en fin de ligne. On peut ainsi visualiser toutes les formes d'un même lemme. Des procédures rapides par simple clic permettent d'affecter à chaque lemme toutes ses formes attestées dans le corpus et de réunir leurs occurrences dans un fichier d'index spécifique.
Il apparaît alors de « vrais » homonymes : après indexation automatique, il est possible de distinguer très rapidement ces homonymes par *« glisser-déposer »* pour éviter ensuite toute confusion[2].

1.2.4. Autres recherches

Au fur et à mesure des utilisations, d'autres recherches plus complexes sont apparues comme nécessaires. Elles ont pu être insérées à l'application.

* Synthèmes ou lexies

L'indexation automatique repose sur la reconnaissance des séparateurs de mots ou l'identification des mots composés à partir de leur tiret. Mais il existe des groupes de mots figés ou des expressions qui ne pouvaient être identifiés comme tels à partir des séparateurs de mots. Ces expressions peuvent être recherchées puis *Corputex* en mémorise les références dans un fichier d'index.

* Recherche de passages à partir de citations

Parfois on peut avoir rencontré une citation du texte sans savoir où elle se trouve dans le corpus : *Corputex* sait la retrouver, pour peu qu'elle soit incluse dans une seule ligne[3].

* Indexation thématique

Un des derniers outils créé dans *Corputex* est la constitution possible, à partir d'une série de chaînes de caractères, d'une fiche d'index réunissant toutes les occurrences synonymes ou coréférentielles d'un thème particulier ou d'un personnage. Cela permet en particulier de pouvoir mémoriser manuellement les occurrences de certaines désignations pronominales d'un

[1] Un instrument de recherche sur corpus comme *Frantex* avait déjà, bien sûr, cette fonction.
[2] Le rythme moyen du traitement de ces homonymes est de l'ordre de 500 occurrences à l'heure. Autrement dit, seul le traitement qui différencie l'article défini « le » du pronom personnel homonyme représente un travail un peu trop long. Mais il est possible de le faire en plusieurs fois. Et à tout moment une erreur peut être corrigée.
La mémorisation d'une catégorisation grammaticale pour chaque lemme permettrait en fait de préparer la répartition en utilisant les tests distributionnels connus : « le » article avec un nom à droite, « le» particule préverbale devant le verbe ou d'autre particules.
[3] En fait il peut aussi parfois reconnaître une « suite » située en début de ligne suivante, mais l'algorithme n'est pas encore sûr à 100%.

personnage ou de certaines périphrases le désignant. L'indexation de toutes ces occurrences dans une fiche spécifique permet de pouvoir retrouver, avec un peu de patience, l'ensemble du parcours d'un personnage dans le corpus. On peut aussi réunir toutes les occurrences rendant compte du parcours thématique d'une notion ou d'un thème, quelles que soient ses différentes désignations[1].

1.2.5. Classer les trouvailles pour pouvoir les différencier et les retrouver facilement

Un fois ces recherches faites, il faut enregistrer le *sous-corpus* considéré, composé des passages intéressant le problème à étudier. *Corputex* permet de classer dans un dossier spécifique les trouvailles ainsi faites, éditées sous formes de fiches qui peuvent, elles-mêmes, recevoir un classement et tous les commentaires utiles. Si, par la suite, le texte du corpus est amélioré, chaque extrait reprendra le nouveau texte pour la bonne raison que les fiches d'extraits ne mémorisent que la référence du passage : elles recopient systématiquement le texte du corpus d'origine lors de l'affichage de la fiche[2].

Le travail d'étude des extraits dans les dossiers consiste d'abord à les classer selon des critères hiérarchisés posés par l'utilisateur. Ce travail de classement est assez rapide, du fait de l'interface de travail[3]. Chaque dossier peut être dupliqué pour être soumis à plusieurs classements successifs qui peuvent, d'ailleurs, relever d'analystes différents. Il est possible aussi de déplacer certains extraits d'un dossier à l'autre soit de façon individuelle, soit de façon collective. Un export de toutes les fiches se fait en format texte.

Chaque extrait, chaque dossier peuvent recevoir des commentaires.

1.2.6. Pouvoir lire tout le cotexte *nécessaire et surtout pouvoir le lire immédiatement !*

Avant d'analyser le sous-corpus, l'utilisateur choisit le nombre de phrases ou de lignes demandées avant et après la ligne référencée. S'il a besoin d'en voir plus, le passage concerné peut être élargi par un simple clic. S'il veut avoir accès à tout le corpus, cela est possible de façon immédiate, en pouvant retrouver, directement sur le corpus, le passage considéré.

Lors de ces lectures visant à mieux identifier le *cotexte* et son *contexte référentiel*, *Corputex* permet de faire toute une série d'investigations sur l'ensemble du corpus, en particulier, d'installer des index thématiques pour mieux se représenter tout les détails utiles de l'intrigue, comme par exemple, ce que vient de vivre le personnage.

Nous remarquerons ici un inconvénient pratique de *Corputex* dont nous reparlerons en fin d'exposé : quand l'utilisateur se met à annoter le corpus, la tâche devient une sorte de tonneau sans fond. Sa curiosité le conduit à faire de nombreuses recherches et comme le corpus est lu comme un rouleau continu et non page par page, le pauvre lecteur finit par ne plus savoir où il est, où il en est et ce dont il s'occupait au départ. S'il utilise l'application pour préparer des cours, surtout s'il se lance à l'improviste dans des recherches non prévues, il finit vite par ne plus avoir de temps de faire le travail projeté au départ.

2. Les conditions théoriques de cette élaboration

Il est sûr que ce serait bien plus simple d'avoir un instrument de recherche ou d'évaluation qui sélectionnerait, sur des critères quantitatifs, ou par analyse automatique, ce qui serait plus pertinent.

Cela réduirait d'autant le fastidieux travail d'analyse extrait après extrait[4]. La productivité du travail de recherche en serait grandement améliorée.

Il faut donc examiner maintenant pourquoi il faut tant tenir à un outil qui oblige à cet énorme travail personnel du chercheur qui, comme le Saint-Thomas des Évangiles, ne voudra prendre comme

---

[1] Il est bien sûr toujours possible de supprimer toutes les références qui s'avéreraient erronées après vérification.

[2] Ce principe a été un peu réduit dans la mesure où il s'est avéré nécessaire de conserver un certain balisage qui met entre chevrons l'empan de texte qui semble pertinent pour attester du fonctionnement de l'extrait par rapport au problème posé. Le texte de la ligne convoquée ou du passage délimité par les chevrons est conservé dans la sous-fiche. Mais une procédure simple – un effacement du champ contenant ce texte – permet de retrouver le texte modifié du corpus.

[3] Si la lecture n'est pas trop difficile on peut atteindre les 100 extraits à l'heure.

[4] Pour permettre un gain de temps, il est cependant possible de faire des sous-sélections au hasard dans le dossier d'extraits. Chaque dossier est pratiquement limité à 200 occurrences.

fait empirique que ce qu'il aura vu et qu'il aura pu interpréter et évaluer. Cette méthodologie s'appuie sur certains principes avancés par la *sémantique interprétative*. Mais nous verrons que le point de vue *praxématique* que j'adopte dans ma démarche en sciences du langage l'exige plus encore.

2.1. Herméneutique philologique

Nous ne reprendrons pas ici tout ce que François Rastier expose très clairement dans *Arts et science du texte,* notamment dans son chapitre III « Philologie numérique »[1]. Ce sont des bases méthodologiques sur lesquelles il n'est pas nécessaire de revenir ici.
Mais pour bien caractériser la méthodologie scientifique qui sous-tend l'écriture de l'application, il faut quand même expliciter quelques éléments que nous formulerons à notre façon, marquant ainsi certaines originalités par rapport au point de vue de la sémantique interprétative.

2.1.1. L'établissement du signifiant

Un helléniste ne peut déroger à un principe philologique fondamental : il faut essayer d'avoir le meilleur texte possible, c'est à dire celui qui a le plus de chances de correspondre à celui qui a été établi par l'auteur. La paléographie ancienne puis les règles de l'édition critique moderne ont établi des critères précis pour atteindre au moins mal cet objectif. Et beaucoup d'éditions modernes montrent de grandes qualités dans ce domaine.
Pour des raisons de temps, ou tout simplement de droits, il n'est pas toujours possible de disposer de la meilleure version numérique du texte étudié. En effet les documents dont on peut disposer sur *Internet* sont, on le sait, des versions souvent fautives dans le détail, du fait qu'elles ont été établies par des logiciels de reconnaissance de caractères et que les relectures qui ont suivi ne peuvent pas ne pas laisser échapper certaines des erreurs de détail que commettent les logiciels. C'est pourquoi on peut découvrir des coquilles. Il faut bien sûr les corriger dès qu'on les repère. Cette possibilité est prévue par *Corputex.*
Si l'on fait le travail de numérisation soi-même à partir d'une édition reconnue (ne serait-ce que celle qui est préconisée par le programme d'agrégation), il faut utiliser une technique qui reproduise exactement la pagination et les retours lignes, autrement dit la *linéation* de l'original afin de pouvoir retrouver très facilement le passage sur l'édition papier[2]. *Corputex* mémorise toutes ces références, y compris les numéros de page, pour permettre toutes les vérifications nécessaires sur l'édition originale. Un système de notes de bas de page permet même de montrer un apparat critique, quand il a été numérisé à partir de l'édition[3].

2.1.2. Le signifiant comme interprétant

À l'inverse de toute une tradition idéaliste qui tend à valoriser le sémantisme et le contenu intellectuel, nous poserons comme axiome d'une herméneutique matérielle qu'un des interprétants fondamentaux pour toute interprétation reste les différences et les réseaux explicités par le signifiant[4]. Il faut donc que celui-ci soit établi avec précision. De plus, l'objectif de fonder des sélections d'énoncés comparables dans une étude en sciences du langage sur des recherches de

---

[1] Ouvr. cit. pp 73 ss

[2] Peut-on garder l'espoir que la possession de l'édition papier autorise le fac simile privé que constitue la numérisation pour satisfaire aux droits des auteurs ? Il y a lieu de penser en tout cas que ce serait une civilité indispensable pour rétribuer le travail des éditeurs.

[3] Toute modification personnelle d'importance peut être saisie dans la même note.
Signalons ici qu'un système de macros pour le logiciel Microsoft Word permet de préparer assez facilement le balisage des références pour le corpus, pour peu que l'on ait eu soin de bien numériser les numéros de page.

[4] La perception de ce signifiant ne peut être indépendante de la reconnaissance de ce signifiant comme produit d'une production de sens. Comme l'écrit François Rastier, « les relations constituantes du sens vont de signifiés en signifiés, mais aussi des signifiés vers les signifiants : ainsi, la *sémiosis* se définit comme un réseau des relations entre signifiés au sein du texte, en considérant les signifiants comme des *interprétants* qui permettent de construire certaines de ces relations (cf. *supra,* chap. I). (…) En d'autres termes, le sens n'est pas donné par un codage préalable qui associerait strictement un signifiant et un signifié ou une classe de signifiés (car la langue n'est pas une nomenclature) : il est produit dans des parcours qui discrétisent et unissent des signifiés entre eux, en passant par des signifiants. » *ouvr. cit*. pp. 103-104. Mais dans cette dialectique, nous insisterons sur le caractère décisif de l'ancrage dans le signifiant concret de tout ce qui contribue à en enrichir la signifiance.

chaînes signifiantes, comme le font *Frantex* et d'autres logiciels comme *Lexis*, *Pandora* et *Corputex,* n'est pas un bricolage provisoire en attendant que l'intelligence artificielle nous donne les instruments pour faire mieux. Il s'agit d'un outil essentiel.

2.1.3. L'activité du linguistique est une démarche herméneutique

François Rastier remarque que toute analyse de propriétés morphosyntaxiques, de significations ou de représentations par le langage est une interprétation et non la simple description de faits de langues, indépendants des cultures qui les ont créées historiquement. Les sciences du langage ne peuvent élaborer que des constructions herméneutiques.
Comment donc est-il possible d'assurer philologiquement de telles constructions, si le linguiste lui-même ne s'emploie pas à vérifier chaque cas avec son propre jugement ? C'est lui qui doit s'assurer de l'adéquation de l'interprétation. Il ne peut laisser cette vérification à la machine. Si donc l'activité herméneutique veut être philologique – c'est à dire aussi soucieuse que possible de garantir l'altérité du texte dans l'intégrité de son signifiant, comme dans la richesse de sa portée signifiante – il faut que l'interprète mette son nez partout pour assumer et vérifier toutes les analyses que le logiciel a mémorisées.

2.2. Un point de vue praxématique

La mention faite ici de la *sémantique interprétative* ne doit pas dissimuler que, dans tous mes travaux en sciences du langage, j'ai toujours pris en compte, y compris comme sémioticien, le point de vue *psychomécanique* et *praxématique*.
Car pendant les dix ans que j'ai passés à Montpellier, j'ai eu l'occasion de travailler avec Robert Lafont et son équipe. Et j'ai trouvé, dans cette théorie matérialiste, une base très solide pour décrire le fonctionnement concret de la langue[1].

* Robert Lafont

Robert Lafont propose une théorie du signe beaucoup plus radicale encore que la sémantique interprétative : pour lui, il n'existe plus de signe biface mais seulement des outils de production de sens, les *praxèmes,* qui sont de simples signifiants liés à une *praxis*, c'est à dire à un programme de sens expérimenté par chaque individu. Ce programme relève plus d'une sorte de mode d'emploi pour un parcours symbolique à travers un lexique virtuellement hiérarchisé que des contenus notionnels à proprement parler.
Il me semble d'ailleurs que l'on devrait remarquer, pour mieux la comprendre, que cette théorie particulière doit plus qu'elle ne le dit à la tradition du refus de la prise en compte du sens défendu par *l'antimentalisme* de la tradition distributionnaliste. Celui-ci refuse de parler du sens et ne s'occupe que de l'analyse des distributions des signifiants : le *praxème,* lui aussi, n'est qu'un signifiant qui sert *d'outil de production de sens* dans le cadre du réglage taxinomique qui définit ses conditions d'emploi. Mais, au-delà de cette tradition, bien sûr, la *praxématique* ne se contente pas de faire confiance au dispositif syntaxique pour expliciter une signifiance, elle ajoute que le sens lui-même est produit par la sélection des signifiants concrètement effectuée par l'énonciateur. Celle-ci consiste en une *visée lexicale*, en un parcours présupposé du locuteur à travers l'organisation hiérarchisée des signifiants du lexique, la *logosphère*. Cette *visée* conduit, par pesée des différences entre les potentialités de glose autrement dit les *traits* affectés à chacun des signifiants, au choix, à une *saisie* du praxème le plus pertinent.
On voit bien ici, qu'au-delà de ce refus de la correspondance biunivoque signifiant/signifié, réifiée dans l'oukase d'une mise en miroir essentialisante, la *praxématique* récupère à sa façon les théories guillaumiennes.

* La psychomécanique comme théorisation d'une herméneutique généralisée

Elle reprend, en particulier, la notion de *visée* et de *saisie* propres à la *psychomécanique guillaumienne*, pour interpréter certains constituants de l'énoncé comme *l'ancrage*, la manifestation des opérations qui ont conduit à leur choix. Dans ces cas, le signifiant mémorise, en quelque sorte,

---

[1] Ce point de vue spécifique nous écartera parfois de la sémantique interprétative qui tend à valoriser le signifié, ce qui est attendu d'une sémantique.
Mais il y a toujours eu une certaine connivence entre François Rastier et la praxématique. Je n'en veux pour preuve que le fait que nous ayons pu organiser ensemble, avec l'équipe praxiling, une série de conférence de François à Montpellier en 1993.

la *praxis* interne, le processus psychique concret qui en a construit le sens. Pour reprendre ce que nous venons d'exposer à propos de la sélection lexicale, la *psychomécanique* du choix d'un lexème le conçoit comme un parcours à travers des signifiants possibles qui inclurait progressivement des traits sémantiques de plus en plus spécificateurs, du fait des différences qu'ils potentialisent, tout en excluant les autres. Si un locuteur dit « vache », cela présuppose qu'il a parcouru l'ensemble du *taxème* des animaux de la ferme pour ne pas choisir, par exemple, le *signifiant* « animal » ni le signifiant « cheval » mais le signifiant « vache»[1]. Selon le point de vue *praxématique,* le fait que l'on puisse analyser le sémantisme de « vache » en traits successifs repose sur une mémoire des possibilités du parcours de visée ; il ne suppose pas, encore une fois, un *signifié* essentiel à ce *praxème*, ce qui permet de comprendre les variations de sens possibles[2].
De ce fait, on peut analyser que l'acte psychomécanique qui conduit à l'emploi de tel *lexème* plutôt que tel autre est une *opération herméneutique*, en ceci qu'il relève d'une *praxis d'interprétation du monde*. C'est d'ailleurs pourquoi, sans nier qu'il y a bien un réel distinct de nous, nous ne pourrons admettre que la *signification* d'un énoncé puisse correspondre de façon totale au référent de réalité qu'il prétend reproduire. Toute mise en référence, toute construction d'une *impression référentielle* est inévitablement abstraction et interprétation. C'est un produit culturel imaginé en autonomie, en distance par rapport à cette réalité. Dès que, petit bébé, nous nous trouvons en train de montrer du doigt quelque chose que nous désirons, ce geste symbolique que nous faisons interprète ce qu'il montre comme désirable. L'objet désigné n'est déjà donc plus une simple réalité. Il s'agit d'une symbolisation de celle-ci qui explicite le désir qu'elle suscite.

* Le X– (lire « X tiret »)

L'interprétation consciente de ces énoncés suppose deux étapes : nous percevons d'abord intuitivement un effet de sens, puis nous verbalisons celui-ci à partir des réglages présupposés de la *praxis* qui ont conduit à le produire.
Après l'avoir soumise à l'équipe *Praxiling* en 1992, j'avais présenté, il y a plus de 10 ans, en ce même lieu, lors d'un colloque de sémiotique[3], une formulation que je continue à défendre, même si elle ne semble pas encore avoir vraiment trouvé d'échos. Nous pouvons observer que tout énoncé perçu ne peut être identifié dans son contenu que par une glose qui l'analyse de la même façon que tout acte langagier encyclopédique analyse une référence au monde. Autrement dit, si nous nous trouvons devant le texte entier de la *Princesse de Clèves* de Madame de Lafayette, tout ce que nous pouvons dire de ce que nous en avons lu est globalement un X– (lire X tiret). C'est un X parce que c'est une inconnue, un *in posse* dont nous ne pouvons prendre conscience qu'en paraphrasant le contenu. Mais c'est un X qui est nécessairement lié à un tiret. Ce tiret est, en fait, une forme de pesée de nécessité, une loi de signifiance qui nous donne l'aptitude à percevoir si ce que nous disons de l'énoncé considéré est conforme ou pas à la signification de celui-ci.

* Les pesées herméneutiques

Au-delà de cette nécessité d'interpréter pour comprendre ce qu'on a compris, il faut constater aussi que cette interprétation repose, comme toutes les autres, sur un complexe d'évaluations intuitives. Pour juger si notre glose élucidant l'inconnue X est conforme au *texte*, que nous définissons comme l'équivalent du tiret, c'est à dire comme ce *poids* complexe de contraintes diverses qui nous conduisent à dire la signification de façon fidèle, nous recourrons à une autre évaluation tout aussi complexe, qui ne peut jamais relever d'une certitude objectivable. Nous posons donc alors toute une série de jugements qui sont du même type que les *jugements de grammaticalité* et *d'acceptabilité* supposés par la *grammaire générative*. Comme l'a montré François Rastier, ces jugements relèvent d'évaluations, je dirai de « *pesées* » purement esthétiques, aussi (peu) sûres que l'acte par lequel nous attribuons une bonne note à une

[1] Nous ne reprenons pas l'analyse classique en praxématique qui fait parcourir tout le vocabulaire scientifique pour faire descendre au praxème « vache » à partit du générique « animal » en passant par le praxème « ruminant », car l'emploi du mot « vache » est plus souvent lié à l'expérience de la ferme qu'à la connaissance zoologique des animaux. Cet autre parcours est bien sûr tout aussi possible. Pour comprendre toutes les potentialités de *signifiants praxèmes* qui pourraient être choisis à la place de « vache » il faut regarder qu'elles sont toutes les gloses possibles du *signifiant* « vache » dans l'énoncé considéré, autrement dit toutes les paraphrases qu'on peut en faire pour dire sa *signification*.
[2] C'est une originalité de la *praxématique* par rapport à la *psychomécanique* guillaumienne qui intègre sans broncher le *signe saussurien*.
[3] voir SADOULET Pierre, 1998.

dissertation qui nous est apparue très brillante, ou que le jugement qui nous conduit au choix d'une pièce de vêtement. Comme dit le proverbe : « Des goûts et des couleurs »…
Pourtant toutes les sciences du langage reposent empiriquement sur ce type de jugement, car elles ne peuvent produire que des gloses, elles-mêmes contrôlées par ces pesées de type esthétique : les « faits de langue » n'existent pas indépendamment de leurs interprétations et d'un jugement sur l'adéquation de ces interprétations.

* Les sciences du langage pour leur vertu heuristique

En outre, lorsqu'on fait de la linguistique, tout changement dans la description généralement admise n'a aucun intérêt, s'il ne permet pas de faire découvrir de nouveaux réglages du signifiant. Si l'on veut que les sciences du langage dépassent quelque peu les œillères de la vieille grammaire, il faut qu'elles confirment leur vertu heuristique. La description la plus utile sera celle qui mettra en conscience une interprétation évidente, un trait du X– , mais un trait que personne jusqu'ici ne savait décrire, c'est à dire expliciter par des mots. C'est le fond de notre métier.

* Les analyses de statistiques lexicales

Si nous laissons aux méthodes quantitatives la fonction d'évaluer ce qui serait un élément saillant, du fait de sa plus grande fréquence, nous risquons prendre en compte un fait différentiel qui ne pourrait relever, par lui-même, d'aucune *praxis* langagière. Du point de vue *praxématique*, une plus grande fréquence ne veut donc rien dire a priori, tant qu'elle n'est pas rapportée à une opération *psychomécanique* : elle décrit des phénomènes qui peuvent tout aussi bien relever du hasard, surtout si l'on mesure bien la petitesse des différences que la quantification peut identifier dans le domaine langagier[1].

Admettons quand même qu'il y ait bien, dans ces particularités retrouvées du signifiant textuel, des phénomènes quantitatifs incontestables en eux-mêmes. Quelle peut être leur pertinence ?
Car donner directement foi à ces données quantitatives, ce serait attribuer aux phénomènes statistiques la capacité de manifester des différences significatives sans l'intervention du moindre sujet humain.
C'est là tout le problème. Ces données quantitatives ne peuvent être pertinentes que si l'on retrouve l'activité de production de sens qui les ont produites. Si ces écarts quantitatifs peuvent être corrélés à un réglage plausible de la *praxis*, la démarche devient acceptable. Tous ces constats quantitatifs doivent donc être suspectés a priori comme illusoires, tant qu'ils ne sont pas rapportés, grâce à une interprétation, à une pratique culturelle de production de sens.
Nous constaterons ainsi que les analyses stylistiques ont souvent recours à l'observation de la fréquence particulière d'un thème, d'un sème ou de tel ou tel morphème grammatical dans l'extrait considéré. Mais ce constat est fait dans une séquence restreinte et en prenant en compte le relevé précis des occurrences et les effets de sens créés par leur récurrence. Rien n'interdit alors d'interpréter ces phénomènes du signifiant, pour peu qu'ils présupposent la *praxis énonciative* qui a convoqué plus souvent ces traits sémantiques, à ce moment de l'énoncé. L'écart de fréquence, qu'on peut identifier comme un phénomène de rythme, devient alors, dans ce cas, et seulement dans ce type de cas, une présomption d'isotopie[2].

* Vive les cas rares sources d'inventions heuristiques

Peu enclins donc à nous laisser séduire par des données quantitatives, nous préférerons jouer plutôt les pêcheurs de perles pour trouver l'occurrence qui nous montrera une production de sens

[1] Les classements fréquentiels, quels qu'il soient, montrent que la plupart des lemmes ne dépassent pas un effectif conséquent dans le corpus donné – le nombre d'occurrences de la plupart des lemmes constitue moins de 0,5% de la masse des occurrences. Les fonctions statistiques qui calculent comment le décompte des attestations d'un lemme dépasse son effectif attendu est une amplification qui dissimule que la valeur attendue est minime donc peu valable sur le plan statistique.

[2] Ce sont d'ailleurs ces phénomènes d'isotopie soit de sèmes soit de rythmes qui sont présentés par François Rastier dans le chapitre « Philologie numérique » d' *Arts et sciences du texte*. En fait, dans les exemples qu'il donne, il semble que les phénomènes quantitatifs ont servis d'indicateurs pour conduire le linguiste à s'interroger sur les réglages et poser une interprétation qui cherche à retrouver l'explication sémantique, le fait de praxis derrière le phénomène statistique.

particulièrement éclairante, celle qui proposera un renouvellement interprétatif, celle qui fera deviner d'autres inventions de langue[1].
Cette position qui peut paraître assez conservatrice, applique assez systématiquement, comme nous l'avons dit, l'attitude d'un Saint-Thomas. Comme praticien d'une herméneutique philologique, nous préférerons examiner, autant que possible, chaque occurrence : nous ne croyons que ce que nous avons vu ou plutôt soupesé et interprété dans la recherche de la glose la plus pertinente et surtout la plus éclairante pour comprendre toute la richesse intersémiotique de la production de sens.

Conclusion : un tonneau sans fond ?

Faute de place, il faut renoncer à raconter en détail comment *Corputex* m'a permis de mener des études relevant des domaines différents des sciences du langage dont j'ai eu à m'occuper : la *morphosyntaxe*, la *lexicologie sémantique* et la *sémiostylistique littéraire* en lien très étroit avec la *sémantique interprétative*. Je ne mentionnerai rapidement qu'un exemple pour chacun de ces domaines.
Une double étude lexicale[2] sur l'adjectif du grec ancien « *axiologos* » et sur le français « *considérable* » repose sur des extraits analysés avec *Corputex*. Le logiciel ne m'a sûrement pas permis de poser le modèle descriptif que j'ai pu proposer, qui repose sur des spéculations sémiotiques et praxématiques, mais il m'a aidé à rassembler les données grecques et françaises. Prenant une posture qui voulait identifier systématiquement les contre-exemples, j'ai été amené à confirmer l'adéquation d'une proposition descriptive qui imaginait un *motif* premier pouvant expliquer la diversité des acceptions[3].
J'ai utilisé aussi *Corputex* pour établir le fonctionnement tensif de toutes les constructions consécutives dans le grec de Strabon et certains textes de romans français du XIXe siècle[4]. Si aucune des constructions corrélatives trouvées ne pouvait mettre en question l'analyse par un sème tensif, absolument évident dans ces cas, je me suis aperçu très vite, par une familiarisation due aux diverses relectures que nous impose l'usage du logiciel, qu'il existait aussi un poids créé par le cotexte précédent qui servait de base d'explication pour maintenir un sème tensif aux adverbes de liaison « si bien que » placés en début de phrase. Ce que représentait le quantificateur « si » dans la locution, c'était tout le poids argumentatif ou affectif de ce qui venait d'être dit.
Enfin *Corputex* m'a servi pour une étude intersémiotique de type thématique, dans un travail sur la conception de la beauté dans *Le Songe de Poliphile* attribué à Francesco Colonna et traduit en français par l'humaniste Jean Martin[5]. Menant une recherche de vocabulaire et une sélection par lecture directe du texte, j'ai pu extraire un sous-corpus que j'ai retravaillé pour choisir finalement les exemples insérés dans l'article. Il faut signaler ici que le travail d'insertion des extraits dans l'article s'avère très rapide, car *Corputex* formate directement les exemples qu'il suffit de recopier via le presse papier de l'ordinateur.
Tous les textes que j'étudie en cours sont, maintenant, systématiquement importés dans une base *Corputex*. Cela me facilite grandement toutes les tâches matérielles nécessaires pour produire les documents pédagogiques, mais surtout, grâce à cet outil, je peux annoter mon corpus et en constituer une fiche de lecture réutilisable par la suite. J'en profite pour « *peser* » dans le texte le fonctionnement des « faits de langue » que j'enseigne, sans chercher à avoir des relevés complets.
Mais cette puissance accrue exige toujours des travaux supplémentaires pour que les résultats soient sûrs, donc exploitables… alors que les rythmes universitaires ne nous laissent plus vraiment le temps de préparer sérieusement nos cours. Il m'est arrivé souvent d'hésiter à me lancer dans la préparation de la version numérique d'une œuvre au programme, car cette préparation du corpus demande un gros travail, qu'il faille numériser le texte à partir de l'édition imprimée ou réaménager

---

[1] voir DELEUZE Gilles, 1993 – *Critique et clinique*, Paris, Minuit. chapitre 1.
[2] Pour l'étude sur le grec voir SADOULET Pierre, 2003.
[3] Sur la notion de *motif* voir ; CADIOT Pierre, VISETTI Yves-Marie, 2001 – *Pour une théorie des formes sémantiques : motifs, profils, thèmes*, Paris : « Formes sémiotiques », PUF.
[4] Voir SADOULET Pierre, 2005.
[5] COLONNA Francesco, Martin Jean tr., 1546 éd. texte italien. 1499 – *Le songe de Poliphile*, Paris, Pour Jacques Kerver aux deux Cochets, Rue St Jaques. Voir l'étude dans SADOULET Pierre, 2003.

une version numérique pour retrouver la pagination voire la linéation de l'édition originale. Sans compter qu'il faut annoter l'œuvre ensuite etc…
Je me suis souvent demandé si, finalement, je n'avais pas conçu, comme je l'ai déjà dit, une application qui fonctionne comme un tonneau des Danaïdes et qui deviendrait, de ce fait, une vraie torture pour l'utilisateur. Car quand on utilise *Corputex*, rien n'est jamais fini, rien n'est définitif et il y a toujours un ouvrage à remettre sur le métier. Mais la richesse de perception sémantique que j'obtiens, je crois, dans mes études, laissent l'impression que l'obligation de relecture créée par *Corputex* permet d'aller beaucoup plus loin dans l'enrichissement de l'interprétation qu'on ne pourrait le faire avec de simples fichiers sur papier qui ne feraient jamais lire autant d'extraits, vu la longueur du travail de copie à la main qu'ils exigent. Ce travail à la main était, lui aussi, finalement, un tonneau sans fond, mais un tout petit tonneau. Avec *Corputex* nous gagnons nettement en puissance et en rapidité.

BIBLIOGRAPHIE

LAFONT, R. 1978. *Le travail et la langue*, Paris, Flammarion.
LAFONT, R. & GARDES-MADRAY F. 1988. *Introduction à l'analyse textuelle réédition 88*, Montpellier, Langue et Praxis, Université Paul Valéry.
LAFONT, R. 1994. *Il y a quelqu'un : La parole et le corps*, Montpellier, Praxiling.
RASTIER, F. 1987. *Sémantique interprétative*, Paris, PUF pr. éd 1986, éd. 1991, éd., Paris, PUF, 1996.
RASTIER, F. 1989. *Sens et textualité*, Paris, Hachette.
RASTIER, F. 1994. Le problème du style pour une sémantique du texte, in P. Cahne & G. Molinié (éds), *Qu'est-ce que le style ?*, pp. 263-282, Paris, PUF.
RASTIER, F. (éd). 1996. *Textes et Sens*, Paris, Didier érudition.
RASTIER, F. 2001. *Arts et sciences du texte*, Paris, PUF.
SADOULET, P. 2005. *Corputex : logiciel d'analyse textuelle et de constitution de dossiers d'extraits* (base de données en 4 D), Saint-Étienne. Version 18.
Études de sémiotiques sur des livres d'artistes :
SADOULET, P. 1998. Rhétorique et épaisseur sémantique, in Actes du colloque d'Albi (GDR de sémiotique) *Sémantique et rhétorique* juillet 1995, Toulouse, Editions Universitaires du Sud, pp. 81-103.
Etudes de linguistiques et sémiotiques sur corpus
SADOULET, P. 1980. *Le principe d'économie dans l'expression* (2 tomes), Thèse de troisième cycle (dir. Michel Casevitz), Université de Lyon II.
SADOULET, P. 1996. Un jeu original sur le signifié du Paon : étude de dix estampes de "Pavo, fragment sur le paon bleu", de François Righi, in M.-L. Honeste, R. Sauter (éds.), *Animots*, Université de St Etienne, CIEREC, pp. 159-179.
SADOULET, P. 1998. Du global au local : effets de sens et corrections d'auteur. Travail à partir du manuscrit de « À chaque pas prenant congé », in J.-Y. Debreuille (dir.), *Un poète dans la classe, Jean-Vincent Verdonnet*, Lyon, PUL, pp. 145-164.
Etudes publiée menées à l'aide du logiciel Corputex sur corpus numérisés
SADOULET, P. 2003. Axiólogos chez Strabon. Essai d'apport sémiotique à l'étude d'une polysémie lexicale, in S. Remi-Giraud & L. Panier (dir.), *La polysémie ou l'empire des sens. Lexique, discours,* représentations, Lyon, PUL, p. 65 ss.
SADOULET, P. 2003. L'émotion esthétique et sa représentation verbale dans le Songe de Poliphile (livre I), in C. Ziberberg & F. Parouty (dir.), *Sémiotique et esthétique*, Limoges, PULIM.
SADOULET, P. 2002. Le corps du voyageur dans la description géographique /Traveller's body in geographic description, in *Symposium* organisé par le bureau de l'association internationale de sémiotique, Université de Lyon II septembre 2002.
SADOULET, P. 2005. Le morphème intensif "hôste" dans la géographie de Strabon : entre corrélation et coordination, Communication au colloque "Subordination et corrélation", Bordeaux 26 et 27 septembre 2002, Saint-Étienne.

PFC-ABIDJAN : CHOIX MÉTHODOLOGIQUES LIÉS À L'EXTENSION D'UN CORPUS

Béatrice Akissi BOUTIN
ERSS – Université Toulouse 2

SOMMAIRE

Résumé : *Le corpus PFC (Phonologie du français contemporain : usages, variétés et structure), conçu essentiellement pour examiner la variation géographique, sociale et stylistique de la phonologie du français, comporte de nombreux débouchés puisque les éléments pris en compte dans l'échantillonnage des locuteurs de chaque point d'enquête satisfont aux exigences actuelles de la sociolinguistique et de la linguistique de terrain, dans tout ce qu'elle peut apporter à la phonologie, à la syntaxe, à la sémantique cognitive, à l'analyse de discours, etc.*
Les enquêtes d'Abidjan et Ouagadougou (2004-2005) ont été les premières à étendre le protocole commun de PFC à l'Afrique ; elles seront bientôt suivies de celles du Sénégal et du Mali. Aucun corpus de référence de cette ampleur n'existe encore pour l'Afrique, et aucun corpus africain n'était encore entré dans une telle étude pan-francophone.
Cette extension pose toutefois les questions de l'universalité des outils (ou de l'unité dudit espace francophone), telles que celles du choix de la transcription en français standard et des normes de codage du schwa, domaine privilégié d'observation de PFC.
Nous verrons dans quelle mesure le corpus PFC – Abidjan s'insère dans les problématiques générales de PFC tout en ouvrant de nouvelles perspectives.

Introduction

La communication porte sur l'encodage de phénomènes spécifiquement oraux du français de Côte d'Ivoire dans le cadre du projet PFC « Phonologie du Français Contemporain ». Après une brève présentation du programme PFC et de son extension aux pays africains francophones, je montre la pertinence d'une transcription orthographique, y compris dans plusieurs cas particuliers comme l'alternance d'énoncés d'une autre langue ou les mots et interjections spécifiquement ivoiriens. Je présente aussi la manière dont nous avons adapté le codage des schwas dans des contextes que ne prévoyaient pas les instructions initiales de PFC.

1. Le programme PFC : protocole, transcription et codages

Le projet PFC vise à illustrer la variation phonologique observée dans le français parlé en France et dans les autres zones francophones. La comparabilité des données est garantie par l'adoption d'un protocole et d'outils d'analyses communs dans tous les points d'enquête.
Le protocole comprend deux lectures et deux entretiens ; il est appliqué à une douzaine de locuteurs liés par réseau social dense ou lâche (L. Milroy 1980), par deux enquêteurs, l'un bien connu dans le réseau, l'autre inconnu et présenté comme un ami. Les tâches enregistrées visent à saisir plusieurs registres distincts de chaque locuteur. La durée totale d'enregistrement de chaque locuteur est de 50 minutes environ.
Par la suite, sont entièrement transcris sous Praat le texte et la liste de mots lus, ainsi que 5 minutes d'entretien guidé et 5 minutes d'entretien libre. La tire de transcription est dupliquée une première fois pour recevoir le codage des schwas et une deuxième pour le codage de la liaison.

2. L'extension de PFC aux pays africains francophones

Le projet PFC s'est voulu, dès sa conception, de portée internationale, mais son extension aux situations africaines le confronte à d'autres dynamiques linguistiques.

En Côte d'Ivoire, le français est en contact avec de nombreuses langues dont la vitalité régresse à peine. Il est cependant le véhiculaire principal et son utilisation s'étend à toutes les situations. Les principaux traits qui montrent le dynamisme de son appropriation communautaire sont sa vernacularisation et sa « nativisation » (R. Chaudenson 2000) en cours. D'autres traits manifestent l'appropriation du français, dans le domaine linguistique comme dans le domaine sociolinguistique : la créativité dans le lexique, la syntaxe et la phonologie, la récurrence de ces spécificités dans tous les registres de la langue, qui vont de la variété académique ivoirienne au nouchi, la prise de conscience de la spécificité des usages, la représentation du caractère homogène et identitaire du français de Côte d'Ivoire. Toutes ces caractéristiques confirment la pertinence de l'extension du programme PFC à la Côte d'Ivoire. On peut ajouter la facilité qu'apporte aux études PFC l'existence d'une norme endogène et de descriptions scientifiques de la variété.

Pour l'ensemble des pays d'Afrique, la réflexion a conduit les chercheurs concernés à valoriser la prise en compte de toutes les zones francophones de façon équivalente et à ne rien ajouter au protocole pour l'instant. En l'occurrence, la situation sociolinguistique de la Côte d'Ivoire permet l'application du protocole sans modification.

3. Transcriptions des corpus oraux du français de Côte d'Ivoire

Les corpus oraux du français de Côte d'Ivoire, particulièrement d'Abidjan, qui existent en Europe ont été transcris en caractères phonétiques[1]. A l'Institut de Linguistique Appliquée de l'université Cocody – Abidjan existent au contraire des corpus oraux de plusieurs variétés du français de Côte d'Ivoire, y compris du nouchi, transcris en orthographe standard[2], ou selon l'usage ivoirien.

L'entrée d'un corpus oral du français de Côte d'Ivoire dans des travaux pan-francophones demande en fait une révision de toutes les questions de transcription. Le problème peut aussi être posé dans l'autre sens (du point de vue africain) : dans quelle mesure les outils et les méthodes de traitement utilisés pour les corpus oraux de variétés euro-américaines de français peuvent-ils être utilisés pour des variétés d'Afrique ?

Mon opinion est que les situations sociolinguistiques sont effectivement différentes ainsi que les domaines de variation, mais qu'un grand nombre de critères peuvent rester les mêmes et acquièrent au contraire toute leur validité en étant confrontés à des situations nouvelles. D'autre part, on peut s'appuyer sur les solutions locales ou régionales, qui ne prétendent pas jusqu'à présent avoir une portée universelle, mais qui pourraient pourtant bien servir pourtant de modèles. L'extension des outils et méthodes à toutes les zones francophones apportera donc, en retour, une réflexion technique et théorique nouvelle sur les corpus oraux.

Dans tous les cas, le choix de l'orthographe standard pour la transcription de l'oral spontané est justifié par un certain nombre de motifs connus, d'ordre éthique, théorique ou purement technique.

Je montrerai comment nous avons choisi de résoudre trois types de problèmes de transcription auxquels nous sommes confrontés : les mots et interjections spécifiques, l'insertion d'énoncés d'une autre langue et les contextes particuliers des codages du schwa

4. Mots et interjections spécifiques

La transcription choisie par PFC est celle de l'orthographe standard du français, que l'on relève dans le matériel lexicographique habituel. Il existe aussi pour notre variété de français des travaux lexicographiques spécifiques, notamment le *Lexique français de Côte d'Ivoire*, de S. Lafage 2003-2004 ou l'*Inventaire des particularités lexicales du français en Afrique noire* (AUPELF-UREF 1988), plus général. Ces travaux n'ont pas de portée normative intentionnelle mais constituent une base commune, d'autant qu'ils attestent normalement l'usage ivoirien (ou les usages africains).

Il est important de considérer que la transcription met sur un même pied des variétés de langue de statuts sociolinguistiques très différents. Alors que la norme du français de France est standardisée, fixée, aménagée, suivie pas à pas depuis longtemps, dans les autres pays

[1] M. Jabet 2005 ; K. Ploog 1999 ; Y. Simard 1998 (Voir Simard 2001) ; J.-M. Lescutier 1985 ; J.-L. Hattiger 1981.

[2] Par exemple : E. Niamien N'Gouan 1997 ; F. Leimdorfer *et al*. 1997 (Voir Leimdorfer *et al*. 2002) ; B.A. Boutin 2002 ; J. Kouadio N'Guessan 2005.

francophones, on en est encore à la question parfois brûlante de la prise en compte des normes endogènes. Les *Etats Généraux de l'Enseignement du Français en Afrique subsaharienne francophone* (Libreville, 17–20 mars 2003) viennent de signer la reconnaissance de leur existence. Une double question éthique se pose si l'on fait se côtoyer dans la même transcription des formes de français standard et des formes écrites de mots n'en faisant pas partie. D'une part tout travail de recherche peut être interprété par le public comme une normalisation des formes. D'autre part, la transcription d'une forme non encore standardisée et une position prise sur sa transcription, inconcevable sans tenir compte des travaux existants déjà.

5. Insertion d'énoncés d'une autre langue

Du fait de la situation plurilingue qui prévaut en Côte d'Ivoire, des insertions d'énoncés plus ou moins longs d'autres langues sont fréquentes et apparaissent effectivement dans les enregistrements. Cela n'empêche pas qu'il est tout à fait naturel de mener une conversation totalement en français (de Côte d'Ivoire), contrairement à ce qui se passe dans les pays voisins. La présence de segments de langues autres que le français pose le problème de la transcription. Les principales langues ouest africaines possèdent actuellement une orthographe officielle sur la base de l'alphabet de Bamako 1966. La seule difficulté est que deux graphèmes n'appartiennent pas à l'alphabet latin mais sont empruntés à l'API : E et . Pour éviter la mixité orthographique (Jacques Durand et Jean-Michel Tarrier, à paraître), j'ai pris le parti de transcrire en caractère SAMPA les énoncés d'une autre langue, par ailleurs rares dans le corpus. La fréquence de ces énoncés dans d'autres points d'enquête nous conduira sans doute à une autre solution.

Un phénomène fréquent est l'insertion du nouchi dans les conversations spontanées. Le nouchi est un argot ivoirien fait d'emprunts lexicaux d'origines diverses qui est usuellement transcris en caractère latins. La seule divergence avec l'orthographe française est alors l'utilisation de l'accent circonflexe pour le [E] : « ê » et le [ ] : « ô ». L'écriture de ces mots ne pose donc pas de problème, dans la mesure où peuvent être suivis les critères déjà retenus pour des langues d'Europe comme le basque (Durand et Tarrier, à paraître). Une fois explicité le choix orthographique opéré par le transcripteur, la lecture est d'autant plus facilitée que « l'alignement de la transcription sur le signal sonore permet [...] d'être informé sur la réalité phonético-phonologique du mot transcrit » (Durand et Tarrier, à paraître).

6. Contextes particuliers des codages du schwa

Bien que les critères de codages[1] aient été prévus au départ pour plusieurs zones francophones, l'extension de PFC à l'Afrique, en l'occurrence à la réalité ivoirienne, montre que le codage existant doit faire aussi l'objet d'une extension. Plusieurs cas peuvent ici illustrer les difficultés et les choix méthodologiques opérés :

- **Certains mots sont trop déformés pour être codés**, tout comme *il y a / avait* ([ja] / [javE]) dans les autres variétés de français. Nous avons ainsi : *parce que* prononcé [pas@] / [pa@] ; *comme ça* prononcé [kO~a~] / [kO~O~] ; *quelque chose* prononcé [kESoz] . Dans le premier cas, le schwa est maintenu mais l'identification de la consonne précédente, demandée par le codage, est impossible. Dans les autres cas, le schwa comme son contexte ont disparu.

- **Certaines consonnes finales peuvent être effacées**, en premier lieu les liquides [r] et [l], mais souvent aussi [k] et [t]. Nous avons ainsi : *bord* prononcé [bO], *colère* prononcé [kolE]. On se trouve alors dans une situation exclue du codage, V@, du fait de l'impossibilité d'un schwa dans une telle position quelle que soit la variété de français actuelle. Dans un corpus où ce cas de réduction est si fréquent, on ne peut pas tout simplement se passer de codage : ce serait fausser les résultats et empêcher la comparaison avec d'autres variétés. Une solution a donc dû être trouvée pour coder le phénomène : le chiffre de 3e position, 5, prévu pour indiquer la réduction d'un groupe consonantique par l'effacement d'une consonne, a été utilisé dans ce cas d'effacement de l'unique consonne. Le codage prend alors place après le dernier son prononcé selon les instructions PFC, c'est-à-dire après la voyelle.

[1] Après chaque occurrence possible d'un schwa apparaissent quatre chiffres. Le premier indique la prononciation ou l'absence de schwa, le second indique la position de la syllabe concernée à l'intérieur du mot, le troisième le contexte de la syllabe précédente et le quatrième le contexte de la syllabe suivante. Je ne discuterai ici que le 3e chiffre.

- **Un groupe consonantique peut être réduit par l'effacement de la première consonne**, surtout s'il s'agit d'un [R] / [r]. Nous avons ainsi : *Marc* prononcé [mak], *regarde* prononcé [regad]. Ce cas n'est pas prévu par PFC. Seule est envisagée la réduction de *ministre* à [minis], ou de *titre* à [tit], codée par le chiffre 5 en 3e position. Le codage doit prendre place après la dernière consonne prononcée, donnant par là une première information sur la ou les consonnes effacées. Ce codage a dû être aménagé différemment pour rendre compte de la situation ivoirienne : le 3e chiffre reste 5 et le codage prend place après les deux consonnes graphiques, et éventuellement après le « e ».
Dans tous ces cas, l'extension du codage existant à la réalité ivoirienne a été réalisé en cohérence avec l'ensemble de la notation. Le même chiffre a pu être conservé pour toutes ces réductions, qui deviennent de ce fait facilement repérables, mais le codage occupe des positions différentes, donnant ainsi des indications sur le type de réduction opérée.

Conclusion

La problématique de PFC réside dans sa visée de comparabilité des résultats. Dans ce cadre, le protocole commun doit être suivi rigoureusement dans tous les points d'enquête et les outils et méthodes doivent permettre de refléter la réalité de toutes les données avec la même limpidité. Chaque extension de l'enquête PFC à une nouvelle zone francophone pose cependant de nouveaux défis au programme. Ceux de la situation ivoirienne sont en partie inconnus dans l'état actuel de l'enquête et, de ce fait, particulièrement intéressants pour la réflexion méthodologique de l'ensemble. Ils font apparaître une démarche empirique – inductive de PFC qui aurait pu rester cachée.
Nous avions déjà montré que les spécificités sociolinguistiques ivoiriennes pouvaient s'insérer dans les variables sociolinguistiques prises en compte par le protocole.
Les premiers résultats du traitement du corpus montrent maintenant que les spécificités linguistiques du français de Côte d'Ivoire peuvent être prises en compte selon les références théoriques minimales de PFC : les sons et les mots spécifiques au français de Côte d'Ivoire peuvent faire l'objet d'une transcription orthographique standard ; les variantes de prononciations et les réductions par rapport au français standard ou à d'autres variétés de français peuvent être codées avec le même système.
L'entrée des variétés africaines de français dans des études d'une telle envergure que celle de PFC représente une avancée considérable pour la recherche en Afrique, mais aussi pour la normalisation et la standardisation de ces variétés.

BIBLIOGRAPHIE

BILGER, M. (éd.). 2000. *Corpus. Méthodologie et applications linguistiques*, Paris, Honoré Champion et Presses Universitaires de Perpignan, 380 p.
BLANCHET, Ph. 2000. *La linguistique de terrain. Méthode et théorie. Une approche ethno-sociolinguistique*, Presses Universitaires de Rennes, 145 p.
BOUTIN, B. A. 2002. *Description de la variation : Etudes transformationnelles des phrases du français de Côte d'Ivoire*, Thèse de doctorat, Université de Grenoble 3, 404 p., Coll. Thèses à la carte, Villeneuve sur Ascq, Presses Universitaires du Septentrion.
BOUTIN, B. A. 2003. Des attitudes envers le français en Afrique : Enquête au sein de professions dont l'outil est le français en Côte d'Ivoire, *Education et Sociétés Plurilingues*, 14, pp. 69-84, Aosta, Italie.
BOUTIN, B. A. 2004. PFC en contexte africain : Pré-enquête en Côte d'Ivoire, *Bulletin PFC*, *4*, Colloque International *Phonologie et phonétique du français : du segmental au prosodique*, Publication de l'ERSS, Toulouse 2, **www.projet-pfc.net/bulletin4/bulletin4.htm**.
DELAIS-ROUSSARIE, E. & DURAND J. (éds). 2003. *Corpus et variation en phonologie du français : méthodes et analyses*, Presses Universitaires du Mirail.
DURAND, J. & TARRIER, J.-M. à paraître, PFC, corpus et systèmes de transcription.
DUMONT, P. 2001. *L'interculturel dans l'espace francophone*, Paris, L'Harmattan, 218 p.
HATTIGER, J.-L. 1981. *Morpho-syntaxe du groupe nominal dans un corpus de français populaire d'Abidjan*, Thèse de 3ème cycle, Université de Strasbourg.
Inventaire des particularités lexicales du français en Afrique noire, 1988, UREF, 443 p., Paris, EDICEF / AUPELF.

JABET, M. 2005. *Omission de l'article et du pronom sujet dans le français abidjanais*, Thèse de doctorat, Université de Lund, 205 p.
KOUADIO N'GUESSAN, J. 2005 (à paraître). Le nouchi et les rapports dioula / français, *Des inventaires lexicaux du français en Afrique à la sociologie urbaine … Hommage à Suzanne Lafage*, *Le français en Afrique, Revue du Réseau des Observatoires du Français Contemporain en Afrique*, 19, Paris, Didier- Erudition.
LAFAGE, S. 2003 et 2004. *Le lexique français de Côte d'Ivoire, appropriation et créativité*, tomes 1 et 2. *Le français en Afrique*, *Revue du Réseau des Observatoires du Français Contemporain en Afrique*, 16 et 17, 865 p., Paris, Didier-Erudition.
LEIMDORFER, F. 2002. L'espace public à Abidjan : individus, acteurs et situations de parole, 40 p. in F. Leimdorfer & A. Marie (éds.), *L'Afrique des citadins (Abidjan, Dakar) sociétés civiles en chantier*, Karthala, 400 p.
LESCUTIER, J.-M. 1985. *Recherche sur le processus de réactivation. Cas singulier d'un idiolecte relevant du français populaire d'Abidjan*, Thèse de doctorat, Université de Nice.
MILROY, L. 1980. *Language and Social Networks*, Oxford, Blackwell.
Rapport Final, UNESCO, Réunion d'un groupe d'experts pour l'unification des alphabets des langues nationales, Bamako, 28/02 – 05/03/1966,
http://www.bisharat.net/Documents/index.html
NIAMIEN, N'G. E. 1997. *Le français parlé dans les gares routières d'Abidjan*, mémoire de maîtrise, Université de Cocody - Abidjan.
PLOOG, K. 1999. *Le premier actant en abidjanais : contribution à la syntaxe du non-standard*, Thèse de doctorat, Université Bordeaux 3.
SIMARD, Y. 2001. Français de Côte d'Ivoire : l'actualisation du nom chez des locuteurs non scolarisés, R. Nicolaï (éd.), *Leçons d'Afrique. Filiations ruptures et reconstitution de langues. Un hommage à Gabriel Manessy*, Louvain / Paris, Peeters, pp. 483-496.

ESSAI DE SYNTHÈSE

Michel BALLABRIGA
Université de Toulouse-Le Mirail, C.P.S.T.

Le consensus est très large, semble-t-il, sur *l'utilité*, à distinguer de l'utilisation, des *bases* numérisées qui permettent d'accéder à d'énormes archives consultables immédiatement, et notamment pour les recherches en Sciences Humaines et Sociales. Des voies s'ouvrent à la philologie numérique et les ressources de l'édition numérique, notamment par la constitution d'*appareils critiques* enrichissant considérablement et renouvelant l'apparat critique, contribueront à l'évolution et à la facilitation des études de toute sorte. C'est un *de facto* des nouvelles technologies qui livrent là une matière, une substance en extension rapide dont on ne peut faire quelque chose que par la création de *corpus* dont la constitution dépend de la tâche projetée et on ne peut faire quelque chose d'un corpus de quelque étendue que par l'entremise d'un *outil logiciel* de traitement, d'interrogation. La *tâche* se profile avec le corpus et les outils, nouvelles entités numériques qui doivent être rendues compatibles afin que le corpus offre *prise* à l'outil. Conformément à l'axiome saussurien - c'est le point de vue qui crée l'objet - le corpus semble pouvoir être défini *a minima* (noyau invariant) comme une collection de textes en fonction d'un objectif, d'une recherche.

On peut craindre une instrumentalisation du corpus (et du texte donc) au service du logiciel ; c'est la question des moyens et des fins, pas nouvelle certes, mais d'un autre poids puisque la *machine scientifique oraculaire* a parlé ; la question de l'ADN textuel et la volonté/possibilité de "faire parler les textes", fort stimulantes et suggestives, peuvent éveiller d'autres échos moins plaisants : "les textes vont se mettre à table" (!). Il faut aussi être prudent sur la question de l'attribution : comment distinguera-t-on l'excellent pastiche d'un écrivain qui a intégré, par "innutrition", la façon, le *ductus* d'un auteur ? C'est toute la question de l'imitation quasi-obligée dans la tradition littéraire. Comment distingue-t-on du non assumé sans analyse linguistique précise (cf. "Brutus est un homme honorable" dans le Jules César de Shakespeare), la polyphonie en général ?

Des questions, (davantage liées à l'ethos de la recherche), peuvent apparaître au moment des *choix* au fil des étapes : constitution du corpus (dépendant de l'objectif d'une tâche), codage, annotations, étiquetage, enrichissement etc. et constitution de logiciels d'interrogation des corpus qui reposent sur certaines postures (théoriques, méthodologiques, épistémologiques) qu'il convient de ne pas *naturaliser* même par la technique ("ça va de soi") car elles ont des incidences sur la recherche et ses résultats : ceux-ci sont bien sûr fonction d'une théorie linguistique (d'une grammaire si on veut, à *évaluer* ainsi que les *résultats*) que l'on peut "critiquer".

On peut avoir l'impression que l'outil est plus complexe que les résultats qui prennent valeur de confirmation du su, du pressenti ; on peut avoir le sentiment que le corpus (global) mange le texte (comme localité) ; même si on fait des retours nécessaires mais ponctuels au texte (ou mieux, contexte), *on peut pratiquer ce genre de recherches sans lire les textes du corpus*, contre-partie de la globalisation (inversement, la brièveté du texte rend ces analyses délicates : ce type de recherche a besoin d'un corpus étendu). Est-ce qu'un échantillonnage ne suffirait pas pour certains discours répétitifs, peu innovants, où le figement et le stéréotype sont importants et presque de mise, pour formuler des interprétations, quitte à confirmer et à affiner par le reste du corpus ? Une bonne connaissance du genre permet peut-être une certaine économie descriptive logicielle. Tous les textes ne sont probablement pas justiciables d'une même méthode.

L'interprétation s'est parfois déplacée : corpus codé----> analyses automatiques----> résultats----> commentaires ; *on interprète des résultats (indirects) produits par l'assistance logicielle*, un autre texte, second ; l'interprétation apparaît aussi entre les résultats et les commentaires. *Cette phase interprétative ne paraît guère spécifiée* ; d'où, peut-être, le désir - scrupule tout à fait honorable - de retarder le plus possible et de contrôler l'interprétation - qui apparaît comme un principe de plaisir - en donnant la part la plus importante à la description - qui participerait du principe de réalité - et la mention (avec quelque regret) du nécessaire saut interprétatif, le plus tard possible... En fait, peut-être que la phase descriptive est plus "confortable", on ne dit pas plus facile, et les principes sont peut-être inversés...

Au cœur des questionnements se trouvent notamment la question de *l'unité* et celle du *changement d'échelle* dont découle la question des rapports entre *micro-analyse et macro-analyse.*

Du type *d'unité* et de sa *catégorisation* - dans les recherches actuelles unités lexicales et parties du discours sont privilégiées dans la conduite de l'interrogation - dépend en grande partie le cadre méthodologique, mais le type d'unité et la catégorisation dépendent de la théorie et de l'épistémologie explicites ou implicites.

Le recours au mot (graphique) comme approximation incontournable est recevable, notamment en l'état actuel des techniques logicielles. La question du figement (de ses degrés) - le mot est une unité figée - et des segments répétés permet de complexifier cette question de *l'unité* (phraséologies, prêt-à-dire relevant de formes de la doxa) et de la *sémiose.* Ces nouvelles méthodes font aussi apercevoir de nouvelles formes insoupçonnées du fait du changement d'échelle et de la rapidité du parcours des associations attestées (en gros le télescope "remplacerait" le microscope) que l'humain seul ne peut réaliser. Ces outils ouvrent incontestablement la voie, entre autres, à une nécessaire *sémantique textuelle historique et comparée*, composante essentielle d'une *sémiotique des cultures*, notamment en permettant des études thématiques et topiques d'envergure, jamais permises jusque là. Il se produit certes une modification importante dans la perception des textes ; de nouveaux objets, de nouvelles unités apparaissent, ainsi que de nouvelles façons d'interpréter. Comme le signalait le texte d'orientation, un nouveau rapport à l'empirique se constitue, susceptible d'entraîner de nouvelles formes d'élaboration des connaissances, mais il convient au plus haut point de ne pas enfermer la démarche dans un seul point de vue méthodologique et théorique, de ne pas le *naturaliser*, et de développer par des voies et instruments spécifiques les aspects plus proprement sémantiques : "L'enjeu consiste à passer du *zero meaning* (chaîne de signifiants avec traitement statistique) à l'analyse thématique, à pallier l'absence de "données sémantiques" en tirant profit de la théorie sémantique" (F. Rastier, 2001, *Arts et Sciences du Texte*, P.U.F., p. 206).

Pour le dire grossièrement, le contenu des mots est constitué de mots et cette configuration interne, sorte de sédimentation au niveau de la langue, est *a priori* variable en contexte du fait d'une dynamique propre de l'échange textuel par actualisation ou inhibition notamment : "L'analyse en traits sémantiques reste cruciale, car deux occurrences d'un topos qui n'ont aucune lexicalisation en commun doivent pouvoir être reconnues ; c'est d'ailleurs le seul moyen de sortir de la logique documentaire du mot-clé" (*ibid*. p. 218)". Ceci nous renvoie aux modes d'existence divers d'une unité : on connaît l'exemple des étudiants pensant avoir entendu le mot *avalanche* dans la phrase : "la neige dévalait furieusement la pente" ; cela est à traiter non pas, de façon réductrice, comme une illusion mais par une approche complexe de la perception sensorielle, mentale et sémantique (cf. aussi *ibid*. p. 200-201 : "alors que ce mot [ennui] se rencontre seulement quatre fois dans Madame Bovary, les composants du thème apparaissent souvent [...] Le mot *ennui* est absent, mais les sèmes caractéristiques du thème de l'Ennui se répètent massivement"). Cela conduit bien sûr aux problèmes philologiques et herméneutiques liés à la numérisation et au traitement informatique et ces phénomènes ne sauraient être évacués (puis niés) parce que la technique actuelle ne sait pas encore les traiter convenablement... L'association de termes sur une *surface* impressionnante et complexe (les *cooccurrents*) ne doit pas masquer la question (perceptive et sémantique) de ce qui se passe, du point de vue de la *profondeur* et du *volume*, sous les mots, dans les mots, entre les mots (les *corrélats*). D'où les travaux (en cours) pour des annotations d'un autre type (plus sémantique) : les isotopies et les thèmes, de nature *sémique*, se distribuent sur des termes qui peuvent être *catégoriellement différents*, deux caractéristiques qui posent problème en lexicométrie. Il est possible que la prise en compte des unités lexicales soit (nécessaire et) suffisante pour certains corpus, qui d'ailleurs peuvent s'accommoder d'un simple balayage. Se pose la question de l'adéquation de la méthode, et de ses phases, à différents discours, ce qui renvoie à leur mode de production. L'analyse en traits sémantiques est-elle à réserver au discours littéraire notamment ? Rien n'est moins sûr : elle est utile pour des discours où le stade lexical *paraît* suffisant, c'est-à-dire en fait qu'on ne peut aller au-delà de ce que permet la méthodologie employée, ce que rendrait possible une analyse qualitative : "le quantitatif et le qualitatif ne s'opposent aucunement: seule une analyse qualitative peut rendre significatifs des phénomènes quantitatifs remarquables" (*ibid*. p. 214). La prise en compte du palier lexical est dans tous les cas nécessaire et il faut probablement prévoir et des *stades* interprétatifs (lexical, sémique, que l'on n'opposera pas, etc.) qui ne sont peut-être pas tous pertinents pour tous

les genres de textes, de corpus, et des *phases* dans la tâche interprétative assistée, où se repose la question des moyens et des fins. De même, l'intérêt des macro-analyses ne périme pas celui des micro-analyses (par exemple sur *un* texte, localité qui peut être considérée comme une globalité) qui demeurent nécessaires dans certaines tâches ou parties de tâches, les résultats des études globales permettant d'ailleurs de mieux fonder et d'enrichir les analyses locales, sans oublier la question du plaisir du texte : le "microscope" demeure donc fort utile.
Le logiciel, qui peut être une fin dans une pratique (création de logiciels), apparaît pour la majorité des pratiques utilisatrices comme un *pouvoir-faire* au sens sémiotique, d'une grande potentialité descriptive. Reste à spécifier ce faire : interprétatif, probatoire, heuristique ; dans ce dernier cas, il peut servir à interroger les textes ; mais de quoi dépend la forme de ces interrogations, qu'est-ce qui les suscite, les oriente ? L'interrogation est au centre d'un faisceau dont les extrémités sont reliées à l'objet, à l'objectif, à la théorie, à l'outil (sans préjuger des relations secondes outil-théorie etc.). Ces faires sont d'ailleurs probablement en interaction. Des choix théoriques président à l'étiquetage, au traitement, à l'analyse, à l'interprétation ; il convient de ne pas en faire des absolus descriptifs, de les naturaliser (même si on reconnaît des variations dans les classifications, mais en restant toutefois dans une théorie), de voir que d'autres descriptions sont possibles à partir d'autres postures théoriques. Ce relativisme, *situé*, s'oppose au dogmatisme et il gagne aussi la notion de corpus qui "présuppose une préconception des applications envisagées" (F. Rastier lors d'une intervention à ce colloque) et induite notamment par des positionnements théoriques et/ou méthodologiques : "L'informatique n'est pas un organon théorique, et son usage ne préjuge en rien le bien-fondé d'une thématique assistée" (*ibid*. p. 214).
Il a été largement question des documents et outils numériques, mais la problématique de l'interprétation demeure centrale et ne saurait se détacher d'abord de certaines questions : qu'est-ce qu'interpréter, qui interprète, quoi interpréter, comment interpréter, pour qui et pour quoi ?
Il convient de ne pas opposer brutalement description et interprétation. Où commence l'interprétation, où s'arrête la description ? Quelle est la part interprétative de la description ? Le *versus* entre ces deux termes n'est peut-être pas de mise dans ce continu à seuil dynamique ; on ne décrit pas sans appuis interprétatifs. Il a été dit dans ce colloque, dans une certaine intention : "la catégorisation, c'est l'interprétation" ; on pourrait aussi bien dire : "la catégorisation, c'est *de* l'interprétation"
Il est patent que l'interprétation - *le fait d'interpréter* - est bien l'affaire de tous. Toutefois, l'interprétation - comme *processus scientifique raisonné* - met en oeuvre une méthodologie appuyée sur une théorie aux bases épistémologiques clarifiées ; étudiant de manière critique les conditions de possibilité de l'interprétation, ses résultats ne sont pas simplement "affaire d'interprétation" au sens obvie. Cet aspect-là, compris ainsi, a été peu présent, sinon représenté.
La considération des *moyens* et des *fins* et de leur *dialectique* semble vraiment cruciale dans la thématique de ce colloque :
- la question des *pratiques sociales* en général paraît primordiale (c'est le "pour qui ?") : si le corpus ainsi que son exploitation dépendent de l'objectif, celui-ci est "validé" par la pratique sociale (ou le niveau de pratique sociale) d'où l'activité part, du fait de quelque initiative, et où elle retourne : les types de corpus sont relatifs aux pratiques sociales et aux tâches ; dans une visée esthétique une œuvre intégrale peut constituer un corpus. Il convient d'avoir des relations indispensables avec les acteurs de la pratique pour ne pas risquer l'involution : la pratique sociale commande. Cette relativisation peut guider la pertinence des tâches entreprises (construction de dictionnaires, grammaires, études de langues de spécialités, études sémantiques...)
- on peut souhaiter une diversification selon les intérêts des utilisateurs et la nature du corpus ; l'instrument informatique, quand il est conçu et utilisé comme outil ou boîte à outils (dans une conception noble du bricolage) serait à modifier en conséquence selon les tâches et parties de tâches - dans des formes d'appropriation, voire de détournement de l'outil - en relation avec une adaptation théorique nécessaire : en changeant de fond applicatif, la théorie comme *forme* peut se *transposer* (cf. la "dégradation" théorique, assez vilainement dénommée).
Enfin, sur cette question du corpus et des documents numériques, le colloque a été un lieu de rencontre, pouvant dynamiser un vrai projet fédérateur indispensable en Lettres et Sciences Sociales. Le *relativisme situé* qui a été évoqué devrait avoir comme pendant des possibilités réelles et urgentes *d'interdisciplinarité* - statut du corpus dans les différentes disciplines ; méthodes et réflexions différentes ; comparaison des points de vue ; partage des sources, croisement des données et mutualisation pour les chercheurs en Sciences Sociales et Humaines - qui/qu'ouvrirait,

constituée d'objectifs communs, une zone *transdisciplinaire* consistante, afin de se donner les moyens de penser, *culturellement*, la complexité culturelle.

Comité d'initiative :

François Rastier (CNRS, Paris),
Michel Ballabriga (CPST, Université de Toulouse 2),
Pierre Marillaud (CALS).

Comité d'organisation :

Carine Duteil-Mougel (ATILF, CNRS),
Baptiste Foulquié (Université de Toulouse 2),
Robert Gauthier (Université de Toulouse 2),
Béatrix Marillaud (CALS)
Céline Poudat (Université d'Orléans).

Avec le soutien de :

Institut Ferdinand de Saussure (France),

Très Grand Equipement ADONIS, UPS-CNRS 2916,

Colloques d'Albi Langages et Signification,

Centre Pluridisciplinaire de Sémiolinguistique Textuelle.