Catégorisation d'un corpus hétérogène de français médiéval

Sophie Prévost, Serge Heiden, Fernande Dupuis

Catégorisation d'un corpus hétérogène de français médiéval

Fernande Dupuis
ATO – UQAM – Montréal – Canada
dupuis.fernande@uqam.ca

Serge Heiden, Sophie Prévost
UMR 8503 – CNRS / ENS de Fontenay/Saint-Cloud – 92211 Saint-Cloud – France
{slh, prevost}@ens-fcl.fr

Abstract

We have undertaken a morpho-syntactic tagging of the 2 millions words of our corpora of medieval texts. The external and internal heterogeneity of the texts make this task a difficult one. As a result, we had to resort to a double strategy.
Since there is actually no tool adapted to our corpora, we had first to rely on a programmable tagger in order to categorize a first text. As a second step, and building on the results obtained with the first text, we produced a tagger based on contextal rule learning. Using this latter tool we subsequently tagged a second, quite "similar" (in terms of external criteria) text. The success rate was 95%. This two-step process was then used once again to tag additional texts.
The next phase will be to evaluate the heterogeneity of texts according to internal criteria. This task involves the measurement of morpho-syntactic and semantic variation in accordance with statistical methods. It will enable us to correlate internal and external heterogeneity in order to elaborate a "fine-grained" typology of texts.

Résumé

Nous avons entrepris l'étiquetage morpho-syntaxique des 2 millions d'occurrences de notre base de textes médiévaux. L'hétérogénéité externe et interne des textes entre eux complexifie la tâche, ce qui nous a conduit à élaborer une double stratégie.
Il n'existe pas actuellement d'outil adapté à notre corpus, d'où le recours, pour catégoriser un premier texte, à un étiqueteur programmable. Dans un second temps, nous avons construit, à partir de ce texte, un étiqueteur travaillant par apprentissage. Il a été utilisé pour étiqueter un texte "proche" (critères externes) du texte d'apprentissage, et nous avons obtenons un taux de réussite de 95%. La double procédure est ensuite réappliquée pour l'étiquetage des autres textes.
Par ailleurs, nous voulons désormais évaluer l'hétérogénéité entre textes selon des critères internes. Pour cela il s'agit de mesurer la variation morpho-syntaxique et sémantique selon des méthodes statistiques. Il s'agira ensuite de corréler hétérogénéité externe et interne afin d'élaborer une typologie fine des textes.

Mots-clés : corpus hétérogène, étiquetage morpho-syntaxique automatique, apprentissage, typologie, diachronie, morpho-syntaxe.

1. Problématique et enjeux

Notre équipe[1] dispose d'une Base de Français Médiéval, regroupant une cinquantaine de textes intégraux (environ 2 millions d'occurrences). Nous avons entrepris de catégoriser ce corpus,

[1] UMR 8503 "analyses de corpus linguistiques, usages et traitements" - CNRS-ENS Fontenay/Saint-Cloud

avec un jeu d'étiquettes, morpho-syntaxiques pour l'instant, syntaxiques (en termes de fonctions) par la suite, dans la perspective d'études morpho-syntaxiques.

Or il s'agit d'un corpus hétérogène d'un point de vue *externe*. Ainsi les textes qui le constituent s'étendent du 10ème au 16ème siècle, ils sont en anglo-normand, en picard, en champenois, en orléanais..., certains sont en prose, d'autres en vers, on trouve des romans, des récits historiques, des livres de coutumes... Cette hétérogénéité externe induit une hétérogénéité *interne* entre les textes, qui se manifeste dans des variations lexicales et morpho-syntaxiques (désinences verbales et nominales, ordre des mots, sous-catégorisation des verbes...), selon des critères non encore formalisés précisément.

L'hétérogénéité de ce corpus accentue les difficultés inhérentes à toute procédure d'étiquetage, ce qui nous a conduits à concevoir une double stratégie que nous présentons ci-dessous. Elle est par ailleurs à l'origine de l'élaboration d'une typologie des textes associant les critères externes et internes.

2. Difficultés et démarche adoptée

Il s'agit d'un corpus de français médiéval, or il n'existe pas à l'heure actuelle d'étiqueteur utilisable pour cette langue : ceux intégrant des dictionnaires et/ou des règles syntaxiques ne sont pas conçus pour un tel état de langue. Il existe par ailleurs des étiqueteurs procédant par apprentissage de règles à partir d'un texte étiqueté puis application de ces règles sur un texte non étiqueté (technique de Brill (Brill 1992)) : cela suppose cependant l'existence d'un texte déjà étiqueté. Face à cela, nous avons décidé de recourir dans un premier temps au logiciel *SATO*, moteur de filtrage et d'étiquetage programmable[2].

La tâche restait cependant difficile, précisément parce qu'il s'agit de langue ancienne : encore au 16ème siècle, un même texte peut présenter des variations graphiques importantes, et la souplesse dans l'ordre des mots, plus grande qu'aujourd'hui, complexifie les stratégies de codification automatique. Par exemple, en français moderne, il est devenu très rare que le sujet soit postposé au verbe. Encore au 16ème siècle, et a fortiori dans les textes plus anciens, c'était beaucoup plus fréquent (les séquences verbe-sujet étaient parfois aussi nombreuses que celles en sujet-verbe). Sachant que, en outre, la déclinaison nominale est déficiente dès le 13ème siècle et devenue quasi inexistante à partir du 15ème, on peut mesurer la difficulté de l'étiquetage syntaxique à venir, et, dans une mesure certes moindre, celle de l'étiquetage morpho-syntaxique.

Par ailleurs, dans la perspective de l'usage d'une technique du type apprentissage/étiquetage (dès lors qu'un premier texte a été étiqueté), ces mêmes phénomènes de variation morpho-syntaxique étaient censés complexifier la tâche : en français moderne, deux textes présentant des critères externes similaires ont une morphologie et une syntaxe relativement proches, alors que cela n'est pas le cas en langue ancienne. Par conséquent, il n'était pas garanti a priori que la projection sur un autre texte, même "proche" selon des critères externes, soit efficace.

Enfin se posait la question du choix du jeu d'étiquettes, et d'une manière particulièrement délicate dans la mesure où le corpus présente une langue en évolution. Par exemple, alors que le français moderne ne connaît qu'une forme contractée (préposition + déterminant défini), l'ancien français en connaît six[3], qui disparaîtront progressivement à partir du moyen français. Inversement, le déterminant "ledit" n'existe pas en ancien français, il se répand à partir du

[2] Logiciel conçu par F. Daoust de l'UQAM à Montréal, http://fable.ato.uqam.ca/login.html

[3] Par exemple : pronom personnel + pronom personnel : *ge* + *le* > *gel* ; adverbe + pronom personnel : *ne* + *le* > *nel* ; conjonction de subordination + pronom personnel : *se* (= "si" moderne) + *il* > *sil*.

moyen français. Or dans la perspective d'un étiquetage fin, une telle forme n'est pas assimilable à un simple déterminant défini. Dans la mesure où nous souhaitions concevoir un jeu d'étiquettes qui prenne en compte les formes qui apparaissent et celles qui disparaissent au cours de la diachronie 10ème -16ème siècle, une réflexion linguistique assez pointue a été nécessaire afin de ne pas omettre de formes potentielles.

Deux autres objectifs ont en outre guidé l'élaboration de ces étiquettes. Le premier a été le souci de proposer un jeu qui, dans sa conception et dans sa terminologie, ne soit pas en complète rupture avec ceux existant actuellement, l'éparpillement en la matière étant déjà important. Par conséquent, nous avons tenté de nous aligner, au moins dans l'esprit, sur le jeu d'étiquettes proposé dans le cadre du projet d'évaluation *GRACE* des étiqueteurs morpho-syntaxiques pour le français[4].

Par ailleurs, si l'utilisation de notre corpus est prioritairement destinée aux linguistes, cela n'exclut pas un panel d'utilisateurs beaucoup plus large, et pas forcément intéressé par de subtiles nuances linguistiques : il fallait donc concevoir un jeu d'étiquettes suffisamment fin, mais en même temps simplifiable.

Face à ces différents objectifs, nous avons finalement élaboré un jeu de 58 étiquettes, selon un principe conceptuel et terminologique de décompositionnalité. Celui-ci a un double avantage : une seule requête sur une étiquette tronquée permet de grouper plusieurs requêtes[5], et cela autorise aussi ceux qui ne souhaitent pas entrer dans la finesse de notre étiquetage à se servir d'un étiquetage plus grossier, voire de se limiter aux classiques parties du discours[6].

3. Résultats / évaluation

3.1. *SATO*

Pour étiqueter le premier texte, *La Mort Artu*, roman en prose du 13ème siècle (environ 100000 occurrences), nous avons donc utilisé le logiciel *SATO*. Nous avons mis au point un jeu d'environ 400 règles d'étiquetage, fondées sur des expressions régulières[7]. Elles associent morphologie (en particulier les désinences) et syntaxe (par exemple le repérage/étiquetage des verbes conjugués s'appuie pour une bonne part sur la position des pronoms personnels), et combinent application en contexte et hors contexte (avec un système d'héritage réciproque possible entre les deux), avec une procédure finale de désambiguïsation[8]. Le point de départ de la procédure, et l'expérience a montré qu'il était bien fondé, est l'étiquetage des verbes

[4] Dans la pratique, il y a des divergences : le jeu d'étiquettes du projet *Grace* prend en compte certaines données linguistiques ou établit certaines distinctions que nous n'avons pas retenues pour diverses raisons. Inversement, la spécificité de notre base textuelle (textes de langues ancienne sur une diachronie de 6 siècles), et le souci d'affiner certaines catégories morpho-syntaxiques (en particulier pour les mots subordonnants) nous a conduits à introduire des étiquettes qui n'ont pas leur équivalent ailleurs.

[5] Par exemple, en faisant une requête sur "d.*", on obtiendra "dd-", "ds-", "di-", "dk-", "dro"..., c'est à dire l'ensemble des déterminants (du fait que l'étiquette de tous les déterminants commence par "d").

[6] Pour une présentation plus détaillée de l'étiquetage de la Base du Français Médiéval, voir http://www.lexico.ens-fcl.fr/catego_bfm.htm.

[7] Ces règles d'étiquetage ne sont pas appliquées une par une, mais regroupées en fichiers d'exécution ("scénarios").

[8] L'intérêt d'un étiquetage hors contexte -*SATO* génère automatiquement un lexique du texte- est double : d'une part, la projection de cet étiquetage au niveau du texte lorsqu'il s'agit d'unités non ambiguës, et, d'autre part, la génération finale d'un dictionnaire, projetable sur le lexique d'un autre texte. Deux remarques s'imposent sur ces deux points. Tout d'abord, une seconde procédure est actuellement à l'étude. Elle consiste en un étiquetage direct dans le texte (en contexte), et suppose donc une désambiguïsation au fur et à mesure. L'étiquetage du lexique (hors contexte) est obtenu par héritage de celui en contexte. Par ailleurs, notons que, quelle que soit la procédure adoptée, la projection (préalable ou en cours de route) d'un dictionnaire est tout à fait possible. *SATO* a néanmoins l'avantage de permettre de procéder même sans dictionnaire, ce qui était notre cas face au premier texte à étiqueter.

conjugués, pôles qui permettent ensuite un repérage/étiquetage plus aisé d'un nombre important de formes appartenant à diverses catégories. Pour l'ensemble des différentes étapes de la procédure, et au sein de celles-ci, l'ordre d'application des régles est évidemment capital, sous peine d'"écraser" certaines formes déjà étiquetées[9]. La tâche la plus complexe a été, on peut s'en douter, l'élaboration des règles de désambiguïsation (par exemple entre les différentes valeurs de "que"[10], ou bien entre pronoms personnels sujets et compléments[11], entre ces derniers et les déterminants définis, sachant que, à l'ambiguïté des formes "le", "la" et "les", il faut ajouter en ancien français la forme "li", aujourd'hui disparue). Dans l'état actuel de nos travaux, il reste quelques formes pour lesquelles nous n'avons pas encore trouvé de règle d'étiquetage suffisamment efficace.

A l'issue de l'application de la procédure, nous avons atteint un pourcentage de réussite d'environ 60%. Nous pensons pouvoir obtenir un meilleur résultat par l'amélioration de certaines règles. Néanmoins, en dépit d'un succès non encore pleinement satisfaisant, la procédure n'en conserve pas moins un grand intérêt à différents égards : d'une part, elle est instructive du point de vue linguistique, dans la mesure où elle oblige à formaliser certaines données linguistiques, tant sur le plan morphologique que syntaxique. Les étiquetages erronnés sont de ce point de vue particulièrement instructifs, puisqu'ils permettent de repérer certaines régularités.

Par ailleurs, *SATO* offre une syntaxe d'écriture à la fois très souple et puissante, et, qui plus est, les bases de règles sont aisément modifiables (face à deux états de langue différents, elles sont en effet à transformer partiellement, en fonction des connaissances linguistiques préalables que nous avons sur ces deux états de langue).

Une fois *La Mort Artu* entièrement étiqueté et vérifié, nous disposions d'un corpus d'apprentissage pour mettre en oeuvre la technique de Brill. Nous avons décidé de valider la procédure sur un texte "proche" (selon des critères externes : même époque, roman en prose, langue dialectalement voisine) : *La Queste del Saint Graal* (environ 120000 occurrences). Nous pensions cependant que, malgré leur proximité, la variabilité morpho-syntaxique et la souplesse de l'ordre des mots feraient difficulté et induiraient beaucoup d'erreurs. Or, contre toute attente, nous avons obtenu un pourcentage de réussite de 95% (une part non négligeable des erreurs étant en outre simplement due aux mauvais traitements causés par certains mots restés en majuscules). Nous présentons ci-dessous la procédure qui a été mise en oeuvre.

3.2. *EF13 : Construction d'un catégoriseur morpho-syntaxique pour le français du 13ème siècle*

Le programme EF13 (Etiqueteur de Français du 13ième siècle) est le résultat d'une expérience de construction d'un catégoriseur morpho-syntaxique basé sur le texte catégorisé *la Mort Artu* à l'aide de la technique d'étiquetage mise au point par Eric Brill (Brill 1992). Selon cette technique, la mise en oeuvre d'un étiqueteur se déroule en deux phases : une première phase construit des bases de connaissances à partir d'un texte d'apprentissage pré-étiqueté, puis la phase d'étiquetage utilise ces bases pour l'étiquetage de nouveaux textes. La construction de ces bases de connaissances repose sur l'application d'un paradigme original d'apprentissage par transformation de règles d'étiquetage, dirigé par la précision de l'étiquetage final. Nous avons utilisé la version 1.14 pour système Unix du logiciel d'E. Brill pour construire les lexiques de

[9] Toutefois, par précaution, de nombreuses règles stipulent que l'étiquetage ne doit avoir lieu que si la forme ne possède pas encore d'étiquette.
[10] "que" peut avoir 10 étiquettes !
[11] Dimension syntaxique que nous avons d'ores et déjà introduite.

formes composant l'ensemble des textes étiquetés et à étiqueter, les bases de règles lexicales (voir figure 1) permettant de décider de l'étiquetage des formes hors lexiques et les bases de règles contextuelles (voir figure 2) permettant d'affiner le pré-étiquetage initial à partir des lexiques et des règles lexicales.

n°	code	Règle lexicale	Score
1	char	Tout mot contenant le caractère `**e**' est à étiqueter **`v`**	1158.9
2	goodleft	Tout mot précédant le mot **et** est à étiqueter **`ncom`**	331.6
3	hassuf	Tout mot ayant le suffixe «**t**» est à étiqueter **`v`**	214.1
4	hassuf	Tout mot ayant le suffixe «**r**» est à étiqueter **`vinf`**	127.9
5	fchar	Tout mot étiqueté **`NN`** contenant le caractère `**a**' est à étiqueter **`v`**	110.5
6	hassuf	Tout mot ayant le suffixe «**é**» est à étiqueter **`vpp`**	90.7
7	hassuf	Tout mot ayant le suffixe «**ment**» est à étiqueter **`adv`**	71.9
8	goodright	Tout mot suivant le mot **grant** est à étiqueter **`ncom`**	68.7
9	hassuf	Tout mot ayant le suffixe «**ee**» est à étiqueter **`vpp`**	63.0
10	deletesuf	Tout mot auquel on pourrait supprimer le suffixe «**s**» est à étiqueter **`ncom`**	60.9
11	fchar	Tout mot étiqueté **`NN`** contenant le caractère `**o**' est à étiqueter **`v`**	40.8
12	fgoodright	Tout mot étiqueté **`ncom`** suivant le mot **si** est à étiqueter **`adjqual`**	39.9
13	goodright	Tout mot suivant le mot **car** est à étiqueter **`adv`**	34.2
14	goodright	Tout mot suivant le mot **son** est à étiqueter **`ncom`**	31.8
15	fhassuf	Tout mot étiqueté **`v`** ayant le suffixe «**re**» est à étiqueter **`vinf`**	29.0

Figure 1
Les 15 premières règles d'étiquetage des formes inconnues du lexique
(sur un total de 146 règles)

Les règles sont ordonnées par score d'application décroissant (plus le score est élevé, plus la règle permet globalement d'étiqueter le texte en accord avec le texte d'origine). Le code *de chaque règle est le préfixe de celui stocké dans la base lexicale (terminologie de E. Brill). Nous donnons l'interprétation de l'ensemble de la règle à la colonne suivante. Légende des étiquettes : v = verbe, ncom = nom commun, vinf = verbe à l'infinitif, vpp = verbe au participe passé, adjqual = adjectif qualificatif, adv = adverbe.*

Précisons que la segmentation du texte en phrases et en unités lexicales (qui portent les étiquettes) est à la charge de l'utilisateur et doit être réalisée en amont de ces outils. Pour réaliser ces segmentations nous avons utilisé les outils de la Textothèque LML (Heiden 1999b). Les outils LML permettent de gérer des corpus textuels encodés en SGML.

Nous avons de même utilisé la boîte à outils du logiciel d'E. Brill pour appliquer les bases obtenues sur le texte *La Queste del Saint Graal* afin de valider le catégoriseur obtenu. Une première estimation du taux d'erreur est de 5,6% pour un taux de décision de 100% (une étiquette par forme) sur les 2000 premières occurrences du texte.

La simplicité de mise en œuvre de la méthode et la qualité des résultats obtenus justifient l'application de cette méthode de construction d'étiqueteurs pour les textes de la Base de Français Médiéval. Nous pensons endiguer la grande variabilité morphologique diachronique de ces textes en profitant du découplage des bases de règles lexicales et contextuelles.

n°	code	Règle contextuelle
1	NEXTTAG	Remplacer l'étiquette **artdef** par **propers2** si l'étiquette suivante est **v**
2	NEXTTAG	Remplacer l'étiquette **prep** par **proadv** si l'étiquette suivante est **v**
3	NEXTTAG	Remplacer l'étiquette **sub** par **propers2** si l'étiquette suivante est **v**
4	PREVTAG	Remplacer l'étiquette **sub** par **prorel** si l'étiquette précédente est **prodem**
5	PREV1OR2TAG	Remplacer l'étiquette **propers1** par **propers2** si l'une des deux étiquettes précédentes est **propers1**
6	PREV1OR2TAG	Remplacer l'étiquette **propers1** par **propers2** si l'une des deux étiquettes précédentes est **prep**
7	PREVTAG	Remplacer l'étiquette **v** par **vpp** si l'étiquette précédente est **v**
8	NEXT1OR2TAG	Remplacer l'étiquette **adjpos** par **propers2** si l'une des deux étiquettes suivantes est **v**
9	NEXTTAG	Remplacer l'étiquette **propers2** par **sub** si l'étiquette suivante est **propers1**
10	PREVTAG	Remplacer l'étiquette **propers1** par **propers2** si l'étiquette précédente est **adv**
11	NEXT1OR2TAG	Remplacer l'étiquette **proind** par **adjind** si l'une des deux étiquettes suivantes est **ncom**
12	NEXTTAG	Remplacer l'étiquette **prodem** par **adjdem** si l'étiquette suivante est **ncom**
13	WDNEXTTAG	Remplacer l'étiquette **adv** par **coo** si le mot est **ne** et l'étiquette suivante est **adv**
14	PREVWD	Remplacer l'étiquette **prep** par **proind** si le mot précédent est **l'**
15	PREV1OR2WD	Remplacer l'étiquette **prep** par **v** si l'un des deux mots précédents est **n'**

Figure 2
Les 15 premières règles d'affinage contextuel de l'étiquetage
(sur un total de 188 règles)

*Les règles n'ont pas de score mais sont ordonnées selon leur ordre d'application. Comme pour le tableau des règles lexicales, nous donnons l'interprétation de l'ensemble de la règle dans la dernière colonne. Par exemple, à la règle 11 «Remplacer l'étiquette **proind** par **adjind** si l'une des deux étiquettes suivantes est **ncom**», il faut lire «Remplacer l'étiquette **proind** (du mot courant) par (l'étiquette) **adjind** si (l'étiquette d'un des deux mots suivants est) **ncom**».*
Légende des étiquettes : prep = préposition, adjpos = adjectif possessif, adjind = adjectif indéfini, propers = pronom personnel, proadv = pronom adverbial, prorel = pronom relatif, prodem = pronom démonstratif, proind = pronom indéfini, coo = conjonction de coordination, sub = conjonction de subordination.

4. Perspectives d'étiquetage et élaboration d'une nouvelle typologie des textes

Nous envisageons à l'heure actuelle de conserver notre double démarche, associant *SATO* pour l'étiquetage de textes "distants" (selon des critères externes) et *EF13* pour celui de textes "proches" (l'étiquetage d'un texte à partir de l'apprentissage d'un texte "distant" est actuellement en cours : sa vérification réserve peut-être d'heureuses surprises).

Par ailleurs, nous avons désormais pour objectif l'élaboration d'une nouvelle typologie des textes de notre corpus. En effet, notre classification actuelle repose sur des critères purement externes, et définis a priori, qui sont décrits dans un en-tête suivant les recommandations de la TEI (*Text Encoding Initiative*) (Dunlop, 1995). Les principaux sont les suivants : vers/prose, date de la composition, date du ou des manuscrit(s), langue et dialecte(s), "type" de texte (poésie, théâtre, roman, nouvelles, récits "historiques", chansons de gestes, vie de saints/récits hagiographiques, fabliaux/fables, traités, livres de coutumes, chartes, miracles...), destinateur, destinataire, public potentiel. D'une part, il s'agit d'affiner cette classification, en introduisant de nouveaux critères, sachant que certains pris en compte pour des textes appartenant à un état de langue contemporain sont, pour des textes anciens, soit peu pertinents (par exemple la distinction écrit/parlé), soit difficiles à documenter (c'est le cas, par exemple, pour obtenir des informations précises sur le destinataire ou le public potentiel).

D'autre part, nous souhaitons désormais mesurer l'hétérogénéité des textes selon des critères internes. Pour cela, plusieurs démarches complémentaires sont menées. La première s'appuie sur l'analyse des erreurs d'étiquetage à partir d'un corpus d'apprentissage. La seconde consiste à mesurer la variation morpho-syntaxique et lexicale (participation du vocabulaire) d'un texte à l'autre. Pour cela nous avons recours au logiciel *Lexploreur*[12], qui met en oeuvre, entre autres, la méthode des spécificités[13]. Par ailleurs, dans le cadre du projet TyPTex[14], il a été réalisé à partir d'un corpus de français moderne étiqueté morpho-syntaxiquement, un marquage de "traits" morpho-syntaxiques et sémantiques (parmi lesquels les catégories, mais aussi les temps, les modes..., des outils textuels comme "il y a"...). Une analyse statistique multidimensionnelle de vecteurs représentants les textes et composés des fréquences de chaque trait y apparaissant (que nous qualifions de dimension ou critère interne) à alors été menée. A l'issue de celle-ci, on obtient un regroupement des textes en fonction de ces données et, dès lors, une nouvelle typologie, différente de la classification induite par les critères externes, mais avec laquelle il est cependant possible (et même souhaitable !) de la mettre en relation. Une démarche similaire est à l'étude pour notre corpus de français médiéval. Elle exige cependant une importante réflexion sur la pertinence des traits à retenir : si les étiquettes morpho-syntaxiques ou les critères externes utilisés pour un corpus de langue ancienne diffèrent de ceux requis pour la langue moderne, il en va évidemment de même pour les "traits linguistiques".

Aussi difficile que soit la tâche, nous estimons nécessaire et fructueuse cette corrélation de critères externes et internes : elle devrait en effet permettre de mettre au jour une typologie bien plus affinée des textes, non encore réalisée, à notre connaissance, sur un corpus de français médiéval.

Références

ADDA G., MARIANI J., PAROUBEK P. and LECOMTE J. (1999). Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morpho-syntaxiques pour le français. In P. Amsili editor, *Actes de TALN'99 (traitement automatique des langues naturelles),* Cargèse : ATALA, pages 15-24.
BIBER D. (1988). *Variation across speech and writing*. Cambridge. Cambridge University Press.
BIBER D. (1989). A typology of English texts. *Linguistics*, vol. (27) : 3-43.
BIBER D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing,* 5 vol. (4) : 257-270.

[12] Logiciel conçu par S. Heiden, voir (Chetouani et al., ce volume) et (Heiden 1999a).
[13] Pour une présentation de ce calcul, voir (Lafon 1984).
[14] Pour une présentation plus détaillée de ce projet, voir (FLEURY et al., ce volume).

BIBER D. (1995). *Dimensions of register variation : a cross-linguistic comparison*. Cambridge. Cambridge University Press.
BRILL E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Computational Language (ACL) Processing*, Trento.
BRONCKART J.-P., BAIN D., SCHNEUWLY B., DAVAUD C. and PASQUIER A. (1985). *Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse*. Lausanne. Delachaux & Niestlé.
CHETOUANI L., HEIDEN S. (ce volume) Sémantique des noms propres, méthode des cooccurrences.
BRONCKART J.-P. (1996). Genre de textes, types de discours et opérations discursives. *Enjeux*, vol. (37-38) : 31-47.
DAOUST F. (1996). *SATO : système d'analyse de texte par ordinateur, manuel de références,version 4.0*. Université du Québec à Montréal.
DUNLOP D. (1995). Practical considerations in the use of TEI headers in large corpora. *Computers and the Humanities*, vol. (29) : 85-98.
FLEURY S., FOLCH H., HABERT B., HEIDEN S., ILLOUZ G. and LAFON P. (1999). Maîtriser les déluges de données hétérogènes. In *Actes de l'atelier thématique TALN 1999 Corpus et traitement automatique des langues : pour une réflexion méthodologique*, Cargèse, juillet 1999, pages 7-46.
FLEURY S., FOLCH H., HABERT B., HEIDEN S., ILLOUZ G., LAFON P. and PREVOST S. (ce volume) Profilage de textes : cadre de travail et expérience.
HEIDEN S. (1999a). *Lexploreur : manuel utilisateur, v2.3*, `http://diderot.lexico.ens-fcl.fr/doc/lexploreur/`, UMR 8503, CNRS/ENS Fontenay/Saint-Cloud.
HEIDEN S. (1999b), «Encodage uniforme et normalisé de corpus : application à l'étude d'un débat parlementaire», Mots, n°60, pages 113-132, Presses de Sciences Po.
Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris, Slatkine-Champion
MARCHELLO-NIZIA C. (1999). *Le français en diachronie*. Paris. Ophrys.