Étiquetage d'un corpus hétérogène de français médiéval : enjeux et modalités

Serge Heiden, Sophie Prévost

ETIQUETAGE d'un CORPUS HETEROGENE de FRANÇAIS MEDIEVAL: ENJEUX et MODALITES

Sophie Prévost et Serge Heiden (CNRS/ENS-LSH Lyon)

Abstract

We have undertaken a morpho-syntactic tagging of the 2.5 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. 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. The correlation of internal and external heterogeneity will enable us to elaborate a "fine-grained" typology of texts.

1. Problématique et enjeux

1.1. Présentation de La Base de Français Médiéval

Notre équipe[1] dispose d'un corpus de français médiéval, la BFM (Base de Français Médiéval), constitué au fil des années, à l'initiative de C. Marchello-Nizia[2].
La BFM comprend actuellement 50 textes intégraux (à deux exceptions près), qui représentent plus de 2,5 millions de mots. Chacun d'eux dispose d'une concordance "brute" (non catégorisée), sur supports papier et informatique, réalisée avec le logiciel *Analyser* [3].
Si les textes s'étendent du 9ème au 16ème siècle, il s'agit néanmoins majoritairement de textes d'ancien français (72%), dans la mesure où cette base s'est voulue complémentaire de celle développée par l'INaLF pour la réalisation du futur Dictionnaire de Moyen Français.

Pour des raisons juridiques que nous espérons régler prochainement, la BFM reste actuellement privée, au sens où elle ne peut faire l'objet d'une diffusion publique large (sur Internet par exemple). Elle est cependant ouverte à tous ceux (enseignants, chercheurs et étudiants) qui souhaitent la consulter[4].

1.2. Enjeux

[1] UMR 8503 "Analyses de corpus linguistiques", CNRS/ENS-LSH Lyon
[2] Professeur ENS-LSH Lyon /IUF
[3] Logiciel développé par P. Bonnefois (IUFM Limoges/UMR 8503)
[4] Pour cela, contacter Céline Guillot : celine.guillot@ens-lsh.fr

Nous avons décidé de stabiliser, au moins provisoirement, l'enrichissement quantitatif de la BFM, au profit de son enrichissement qualitatif, en développant un étiquetage morpho-syntaxique.
Celui-ci nous permettra de rendre compte des évolutions morpho-syntaxiques du français, visée diachronique qui n'exclut nullement de s'intéresser aussi à des "coupes" synchroniques.

1.3. Spécificités du corpus

Il s'agit d'un corpus hétérogène d'un point de vue externe, et cela pour plusieurs raisons.
La première tient à la datation des textes, qui, comme nous l'avons dit, s'étendent du 9ème au 16ème siècle.
Par ailleurs, certains des textes sont en vers, d'autres en prose, les premiers étant les plus nombreux (64%). Cela est d'ailleurs lié à la sur-représentation des textes d'ancien français, puisque la prose ne se développe qu'à partir de la fin du 12ème siècle.
Une autre source d'hétérogénéité réside dans la diversité des genres. En effet, les textes littéraires dominent, en particulier les textes narratifs (textes "romanesques" : "romans", nouvelles, ou non "romanesques" : récits historiques, mémoires, récits hagiographiques...), mais la BFM comprend aussi de la poésie et du théâtre, ainsi que des textes non littéraires (coutumiers, chartes…).
Enfin, les textes se caractérisent par leur diversité dialectale. En effet, la BFM renferme en premier lieu des textes en anglo-normand, mais nous disposons aussi de textes en champenois, en orléanais, en picard...

Cette hétérogénéité externe engendre, on s'en doute, une hétérogénénité interne entre les textes, qui se manifeste par des phénomènes de variation lexicale et morpho-syntaxique (nous y reviendrons plus longuement).
Or, si cette hétérogénéité, tant externe qu'interne, est appréciable en soi car synonyme de richesse et de diversité, elle accentue cependant les difficultés inhérentes à toute procédure d'étiquetage.
Face à cela, d'une part nous avons adopté une double procédure d'étiquetage, et, d'autre part, nous avons commencé à élaborer une typologie des textes (associant critères externes et internes). C'est principalement la première démarche que nous développerons ici.

2. Difficultés et démarche adoptée

2.1. Difficultés préalables

Il n'existe pas à l'heure actuelle d'étiqueteur adapté à notre corpus. En effet, d'une part, on ne trouve pas d'étiqueteur conçu pour le français médiéval intégrant règles de désambiguisation et dictionnaire, et, d'autre part, le recours à un étiqueteur procédant par apprentissage (à partir des règles d'un texte déjà étiqueté puis application de ces règles sur le texte à étiqueter) suppose un texte déjà étiqueté.

Face à cette carence, nous avons adopté une double stratégie. Dans un premier temps, nous avons eu recours à *SATO*[5], moteur de filtrage et d'étiquetage programmable, puis, à partir d'un premier texte étiqueté, nous avons développé une procédure d'apprentissage inspirée de la technique d'E. Brill (Brill 1992).
Cela n'a cependant pas résolu les difficultés propres à notre corpus.

2.2. Hétérogénéité interne entre les textes

Cette hétérogénéité interne se manifeste de plusieurs manières.
Ainsi, entre deux textes, en particulier séparés dans le temps ou appartenant à des dialectes différents, on observe que, outre l'enrichissement lexical naturel, des catégories grammaticales apparaissent et d'autres disparaissent. Par exemple, le déterminant *ledit* (avec toutes ses graphies) n'apparaît qu'en moyen français.
Le phénomène le plus fréquent reste cependant la variation morpho-syntaxique (plusieurs formes pour un même mot).
Cette variation apparaît entre textes répondant à des critères externes différents (en particulier la date et le dialecte), mais aussi entre textes répondant à de mêmes critères externes.
On rencontre par exemple, au sein de la BFM, les formes suivantes pour la 4ème personne de l'imparfait du verbe "avoir" : *aviiens, avïens, aviens, aviions, avyons*, et les formes suivantes pour l'adverbe d'intensité "moult" : *molt, mult, mout, moult*.
Mais, plus déroutant pour le locuteur/lecteur moderne, cette hétérogénéité se rencontre aussi au sein d'un même texte. Ainsi, dans les *Quinze Joyes de Mariage* (début 15ème siècle), on trouve pour l'adverbe "aussi" : *auxi, aussi, aussy*. Dans *Jehan de Paris* (fin 15ème), on a, pour "lesquels" les formes *lesquelz, lesquieulx* et, pour "lesdits" les formes *lesdicts, lesdictz*.
Par ailleurs, les textes présentent des phénomènes d'homonymie, beaucoup plus fréquents qu'en français moderne. Ainsi, jusqu'au milieu du 15ème siècle, la langue ne fait pas de distinction entre *a* verbe et *a* préposition. Autre exemple, dans *La Mort Artu*, (début 13ème), la forme *aus* peut correspondre à un pronom personnel ("eux") ou à un déterminant contracté ("aux"). Et l'on rencontre bien sûr aussi de mêmes phénomènes de polysémie qu'en français moderne.

Enfin, il faut ajouter à cela une souplesse dans l'ordre des mots beaucoup plus grande qu'en français moderne, cela encore au 16ème siècle. Par exemple, de nos jours, le sujet est rarement postverbal dans les propositions déclaratives : c'est en revanche beaucoup plus fréquent en français médiéval. Si l'on considère que la déclinaison, déficiente dès le 13ème siècle, a quasiment disparu au 15ème siècle, on conçoit les difficultés engendrées pour l'étiquetage morpho-syntaxique.

D'une manière générale, ces différents facteurs d'hétérogénéité complexifient notre démarche d'étiquetage.
D'une part, avec *SATO*, ils rendent plus difficiles l'élaboration des règles d'étiquetage automatiques. D'autre part, dans la perspective d'un étiquetage par apprentissage, on n'est plus assuré du succès de la projection de bases de connaissance, même sur un texte proche selon

[5] Logiciel conçu par F. Daoust (UQAM, Montréal), http://fable.ato.uqam.ca/login.html

des critères externes (alors qu'en français moderne, deux textes proches selon des critères externes ont une morpho-syntaxe assez voisine).

Il faut signaler que si cette hétérogénéité n'a pas encore été formalisée ni mesurée rigoureusement, cela constitue justement l'un de nos axes de recherche (cf. la présentation de nos objectifs, *infra*).

Enfin, dernière difficulté, il a fallu décider du choix d'un jeu d'étiquettes.

2.3. Choix d'un jeu d'étiquettes :

La tâche était particulièrement difficile du fait que l'on est face à une langue en évolution : des formes apparaissent tandis que d'autres disparaissent.
Par exemple, le français moderne ne connaît qu'un type de formes contractées : préposition + déterminant (avec des sous-types : préposition + déterminant défini : *des*, *au*… mais aussi préposition + "ledit" : *dudit*… ou préposition + "lequel" : *auquel*…). En ancien français, on rencontre en revanche 6 formes contractées (qui disparaissent progressivement en moyen français), par exemple : pronom personnel + pronom personnel (*ge* + *le* > *gel*), adverbe + pronom personnel (*ne* + *le* > *nel*), conjonction de subordination + pronom personnel (*se*[6] + *il* > *sil*).
Inversement, la forme *ledit* n'existe pas en ancien français, elle n'apparaît qu'en moyen français (et 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 un jeu d'étiquettes incluant toutes les formes de la diachronie 9ème-16ème siècle, il nous a fallu mener une réflexion assez pointue pour ne pas omettre de formes potentielles.

Par ailleurs, deux autres principes ont guidé l'élaboration du jeu d'étiquettes.
Le premier était le désir d'élaborer un jeu d'étiquettes standardisé le plus possible, dans sa conception et sa terminologie, avec ceux utilisés pour le français moderne (l'éparpillement est déjà suffisamment grand en la matière). Il a donc été aligné sur le jeu proposé dans le cadre du projet d'évaluation *Grace* des étiqueteurs morpho-syntaxiques pour le français moderne[7].
Par ailleurs, il s'agissait pour nous d'obtenir un jeu fin mais accessible de manière sous-spécifiée : on a donc opté pour un principe de décompositionnalité des étiquettes (Ada et *alii* 1999).
En effet, d'une part, tout le monde n'est pas intéressé par un étiquetage trop fin (or le principe de décompositionnalité permet de se limiter aux traditionnelles parties du discours), et, d'autre part, une telle démarche permet de grouper plusieurs requêtes en une seule requête sur une étiquette tronquée. Par exemple, si avec le moteur de recherche de *SATO*, on formule une requête sur "d.*", on obtiendra tous les déterminants ("dd-", "ds-", "di-", "dk-"…).

---

[6] *se* = "si" moderne.

[7] Même si, dans la pratique, il y a des divergences : *Grace* prend en compte certaines données linguistiques ou établit certaines distinctions que nous n'avons pas retenues. Inversement, la spécificité de notre corpus (langue ancienne et diachronie de 6 siècles) et le souci d'affiner certaines catégories morpho-syntaxiques (en particulier les subordonnants), nous ont conduits à introduire des étiquettes qui n'ont pas leur équivalent ailleurs.

Le jeu finalement mis au point comporte 60 étiquettes (à partir des 8 catégories du discours)[8].

3. Résultats / Evaluation

3.1. *SATO*

Nous avons donc eu recours à *SATO* pour étiqueter le premier texte, *La Mort Artu*, roman en prose du 13ème siècle (100000 occurrences).
Ce logiciel présente un double avantage : il peut fonctionner sans dictionnaire intégré (mais aussi avec), et il permet par ailleurs la création de ses propres règles d'étiquetage.
Il a ainsi été élaboré 400 règles d'étiquetage, regroupées en une vingtaine de "scénarios" (fichiers d'exécution).
Certaines de ces règles sont applicables en contexte, d'autres hors contexte (au niveau des formes du lexique, lequel est généré automatiquement)[9], et notre stratégie consiste en un va-et-vient entre étiquetage hors contexte (pour les formes non ambiguës) et en contexte, avec un système d'héritage entre les deux[10].
Les règles sont fondées sur des expressions régulières (au sens large du terme), qui associent morphologie (en particulier les désinences) et syntaxe (on s'appuie sur des formes ou sur des valeurs déjà attribuées).

Il est à noter que nous procédons à un étiquetage prioritaire des verbes conjugués. L'expérience nous a en effet prouvé qu'ils constituent des "ilôts de confiance" : leur repérage permet ensuite celui d'un nombre important de formes.
A cet égard, voici un exemple de la démarche démarche adoptée pour l'étiquetage des verbes conjugués :
Un premier scénario, appliqué en contexte, associe désinences finales et repérage des pronoms personnels (compléments et sujets). Les contraintes contextuelles étant suffisamment fortes, aucune vérification n'est nécessaire (taux d'erreur < 0.5%). Les étiquettes sont ensuite transférées au niveau du lexique. Ce scénario n'a cependant pas permis d'étiqueter les verbes sans sujet ou ceux à sujet nominal. Avant l'étape suivante, il est nécessaire de repérer en contexte le maximum de participe passés. En effet, à cette époque, les cas d'ambiguité entre participe et verbes conjugués sont fréquents, en particulier parce que la 5ème personne de plusieurs temps des verbes du 1er groupe peut se terminer en "és", et que, à l'inverse, au masculin pluriel, la désinence finale des participe passés de ce même groupe est en "ez". Après cette étape intermédiaire, un scénario, fondé sur les seules désinences, étiquette hors contexte (au niveau du lexique) des formes verbales non encore étiquetées. Après vérification, les valeurs sont projetées en contexte.

[8] Pour une présentation détaillée de l'étiquetage de la BFM avec explicitation des choix opérés : `http://www.lexico.ens-lsh.fr/catego_bfm.htm`.
[9] L'étiquetage du lexique permet de générer un dictionnaire, ensuite projetable sur un autre texte, que celui-ci soit déjà partiellement étiqueté ou non.
[10] Il reste cependant encore une part -faible- d'étiquetage manuel, et par ailleurs, quelques formes pour lesquelles nous n'avons pas de règles suffisamment efficaces.

D'une manière générale, l'ordre d'application des règles au sein des scénarios et l'ordre d'application des scénarios revêtent une grande importance.
La tâche la plus complexe demeure par ailleurs l'élaboration des règles de désambiguisation. Cette dernière s'avère particulièrement difficile pour distinguer les différents mots subordonnants[11], les différentes valeurs de *que*[12], les pronoms personnels sujets de ceux qui sont compléments[13], et les pronoms personnels compléments des déterminants définis[14].

Voici, à titre d'illustration, les 3 premières règles du scénario de désambiguïsation de *bien* :

1) Si on a : déterminant
+ (facultativement) adjectif qualificatif
+ *bien*
+ une forme qui ne soit ni participe, ni adjectif qualificatif, ni adverbe
alors attribuer la valeur "nom commun" à *bien*

2) Si on a : *bien*, sans étiquette,
+ adjectif qualificatif, participe, adverbe, ou préposition
alors, attribuer la valeur "adverbe" à *bien*

3) Si on a : verbe conjugué + *bien*, sans étiquette
=> attribuer la valeur "adverbe" à *bien*

A ce jour, 4 textes ont été étiquetés intégralement avec *SATO* : 1 texte du 13ème siècle : *La Mort Artu* (100000 occurrences), et 3 textes du 15ème siècle : *Les Quinze joyes de Mariage* (40000), le premier livre des *Mémoires* de Commynes (30000), et *Le roman de Jehan de Paris* (30000).

Le taux de réussite de notre procédure est actuellement de 75%, et nous pensons obtenir un meilleur score par l'amélioration des règles.
Même si le succès n'est pas encore pleinement satisfaisant[15], il reste que la procédure présente un grand intérêt, pour deux raisons majeures. D'une part, elle est instructive du point de vue linguistique, car elle oblige à formaliser certaines données, tant morphologiques que syntaxiques. Les étiquetages erronnés sont à cet égard particulièrement instructifs puisqu'ils permettent de repérer certaines régularités. D'autre part, *SATO* présente une syntaxe d'écriture souple et puissante, les bases de règles étant aisément modifiables (et elles sont partiellement à modifier face à deux états de langue différents).

Une fois un texte étiqueté, on disposait d'un corpus d'apprentissage pour la mise en œuvre de la technique d'étiquetage par apprentissage. On a choisi pour cela le texte le plus important, *la*

[11] Il y en a plus de 10 selon notre jeu d'étiquettes (pronoms, déterminants et conjonctions).
[12] *Que* est en effet subordonnant (conjonction, pronom) mais aussi adverbe (simple ou exclamatif), et, en ancien français, conjonction de coordination.
[13] Pour *nous* et *vous*, d'autant plus que le sujet peut être postverbal.
[14] Sachant que à *le / la / les*, il faut ajouter *li*, et que *la* adverbe ne prend pas d'accent avant le milieu du 15ème siècle.
[15] Il n'est cependant pas si médiocre au vu de la difficulté à étiqueter du français médiéval, et qui plus est, en évolution.

Mort Artu, et décidé de le valider sur un texte "proche" selon des critères externes, c'est à dire en prose, de la même époque, et d'un dialecte proche : *La Queste del Saint Graal* (120000 occurrences).

3.2. Construction d'un catégoriseur par apprentissage : EF 13

La seconde procédure d'étiquetage, développée par S. Heiden dans le cadre de notre laboratoire, a donc consisté en la construction d'un catégoriseur morpho-syntaxique à partir de *la Mort Artu* catégorisé, cela à l'aide de la technique d'étiquetage mise au point par E. Brill (Brill 1992)[16].
Et, malgré la variabilité morpho-syntaxique et la souplesse de l'ordre des mots, nous avons obtenu un pourcentage de réussite de 95% (pour un pourcentage de décision de 100%).

La mise en oeuvre se fait en 2 phases.
Dans un premier temps, il s'agit d'élaborer les bases de connaissances à partir d'un texte d'apprentissage pré-étiqueté. Pour cela, on utilise la version 1.14 pour système Unix du logiciel de Brill, afin de construire les lexiques de formes des textes étiqueté et à étiqueter, les bases de règles lexicales (étiquetage des formes hors lexique), et, enfin, les bases de règles contextuelles (affinement du préétiquetage initial à partir du lexique et des règles lexicales).
Dans un second temps, ces bases sont utilisées pour l'étiquetage d'un nouveau texte.

Voici, pour l'expérience "*Artu-Graal*", les 4 premières règles lexicales d'étiquetage des formes inconnues du lexique (sur un total de 146) :

1	Tout mot contenant le caractère "e" est à étiqueter "verbe conjugué"	1158.9
2	Tout mot précédant le mot "et" est à étiqueter "nom commun"	331.6
3	Tout mot ayant le suffixe "t" est à étiqueter "verbe conjugué"	214.1
4	Tout mot ayant le suffixe "r" est à étiqueter "verbe infinitif"	127.9

(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)

Voici par ailleurs les 4 premières règles d'affinage contextuel de l'étiquetage (sur un total de 188) :

1 Remplacer l'étiquette "déterminant défini" par "pronom personnel complément" si l'étiquette suivante est "verbe conjugué"

2 Remplacer l'étiquette "préposition" par "pronom adverbial" si l'étiquette suivante est "verbe conjugué"

[16] Pour une présentation plus détaillée de la procédure, contacter Serge Heiden : slh@ens-lsh.fr

3 Remplacer l'étiquette "conjonction de subordination" par "pronom personnel complément" si l'étiquette suivante est "verbe conjugué"

4 Remplacer l'étiquette "conjonction de subordination" par "pronom relatif" si l'étiquette précédente est "pronom démonstratif"

(les règles sont ordonnées selon leur ordre d'application)

Il faut préciser que la segmentation du texte en phrases et en unités lexicales (qui portent les étiquettes) est à la charge de l'utilisateur et doit donc être réalisée en amont de ces outils. Dans le cas présent, pour réaliser ces segmentations, on s'est servi des outils de la Textothèque LML (Heiden 1999b), ceux-ci permettant de gérer des corpus textuels encodés en SGML.
On a par ailleurs utilisé la boîte à outils du logiciel d'E. Brill pour l'application des bases obtenues sur *La Queste del Saint Graal*, afin de valider le catégoriseur obtenu.

Plusieurs projets sont en cours dans le cadre de l'étiquetage par apprentissage.
Il s'agit tout d'abord de valider la procédure sur un corpus d'application très proche (selon des critères externes) du corpus d'apprentissage. Disposant du premier livre étiqueté (par *SATO*) des *Mémoires* de Commynes, ce texte peut servir à la catégorisation des livres suivants.
Par ailleurs, une application est prévue à partir d'un corpus d'apprentissage relativement distant du corpus d'application (un siècle ou plus). Enfin, nous avons commencé à élaborer un corpus d'apprentissage associant plusieurs textes appartenant à des époques différentes (ancien et moyen français) afin de le projeter sur un texte d'ancien français et sur un de moyen français.

4. Objectifs

4.1. Etiquetage

Il est actuellement prévu de maintenir la double démarche adoptée jusqu'ici. Plus précisément, on réservera *SATO* pour l'étiquetage des textes distants (selon des critères externes) et la technique d'apprentissage pour les textes proches. Il se peut toutefois que cette dernière s'avère très performante pour l'apprentissage sur un corpus distant du corpus d'application. Elle sera dans ce cas plus largement appliquée.
L'utilisation de *SATO* sera néanmoins conservée, pour deux raisons. La première est que c'est à l'aide de *SATO* que nous procédons à la correction des étiquettes. Par ailleurs, ce logiciel permet un étiquetage partiel des textes, ce que n'autorise pas la procédure par apprentissage, qui catégorise le texte dans son intégralité. Or, dans la mesure où la vérification/correction de l'étiquetage est une tâche assez lourde, et qu'il n'est pas forcément nécessaire, dans le cadre de nos recherches linguistiques actuelles, de disposer de l'ensemble des textes étiquetés à 100%, il est prévu d'associer catégorisations intégrale et partielle, selon les textes.

Si l'étiquetage morpho-syntaxique est fort instructif sur le fonctionnement de la langue, il ne constitue pas pour autant une fin en soi : il se situe en amont de l'analyse linguistique.
Outre les diverses analyses linguistiques que permet ce corpus étiqueté, il s'agit maintenant pour nous d'évaluer l'hétérogénéité entre textes selon des critères internes.

4.2. Elaboration d'une typologie des textes

La classification actuelle des textes de la BFM repose sur des critères externes, définis *a priori*. Il s'agit de la date de composition de l'œuvre ou du document, de la distinction prose/vers, du genre et du dialecte. Ces informations sont décrites dans un cartouche suivant les recommandations de la TEI (Dunlop 1995)[17].

Il s'agit désormais de prendre en compte l'hétérogénéité entre les textes selon des critères internes. Pour cela, plusieurs démarches complémentaires sont prévues.
La première consiste en une analyse des erreurs d'étiquetage à partir d'un corpus d'apprentissage.
La seconde s'appuie sur la mesure de la variation morpho-syntaxique et lexicale d'un texte à l'autre, et l'on utilise pour cela le logiciel *Weblex* développé par S. Heiden (Heiden, 1999a).
Enfin, nous avons en projet la réalisation d'une analyse statistique multidimensionnelle, à partir du marquage de traits linguistiques. Cette démarche s'inspire de celle menée dans le cadre du projet de profilage de textes *TyPTex* (Fleury et *alii* 2000). A partir d'un corpus de français moderne étiqueté morpho-syntaxiquement, il a été réalisé un marquage de "traits" morpho-syntaxiques et sémantiques (catégories, temps, modes..., outils textuels comme *il y a*...). On a ensuite mené une analyse statistique multidimensionnelle de vecteurs représentant les textes et composés des fréquences de chaque trait y apparaissant (dimension/critère interne). Cela a permis l'obtention d'un regroupement des textes en fonction de ces données, et, partant, une nouvelle typologie, différente de la classification induite par les critères externes.
Une démarche similaire est donc en voie de réalisation pour notre corpus de français médiéval. Elle suppose préalablement une importante réflexion sur la pertinence des traits à retenir. En effet, si les étiquettes morpho-syntaxiques et les critères externes utilisés pour un corpus de langue ancienne diffèrent de ceux requis pour la langue moderne, *a fortiori* les "traits linguistiques" ne sont pas les mêmes.

L'ensemble de ces trois démarches devrait permettre la mise au jour d'une typologie affinée des textes, non encore réalisée, à notre connaissance, pour un corpus de français médiéval.

Références bibliographiques :

[17] Il faut noter que certains des critères retenus pour la classification des textes modernes sont peu pertinents pour la langue ancienne (écrit/parlé par exemple), ou difficiles à documenter (informations précises sur le destinataire ou le public potentiel).

Adda, G. / Mariani, J. / Paroubek, P. /Lecomte, J. (1999) : "Métrique et premiers résultats de l'évaluation *GRACE* des étiqueteurs morpho-syntaxiques pour le français", in : Amsili, P. (ed.) : *Actes de TALN'99 (traitement automatique des langues naturelles), Cargèse*, ATALA, 15-24.

Biber, Douglas (1988) : *Variation across speech and writing*, Cambridge : Cambridge University Press.

Biber, Douglas (1989) : "A typology of English texts", *Linguistics* 27, 3-43.

Biber, Douglas (1990) : "Methodological issues regarding corpus-based analyses of linguistic variation", *Literary and Linguistic Computing*, 5 vol. 4, 257-270.

Biber, Douglas (1995) : *Dimensions of register variation : a cross-linguistic comparison*, Cambridge : Cambridge University Press.

Brill, Eric (1992) : "A simple rule-based part of speech tagger", *Proceedings of the Third Conference on Applied Computational Language (ACL) Processing*, Trento.

Bronckard, J.-P. / Bain, D. / Schnneuwly, B. / Davaud, C. / Pasquier, A. (1985) : *Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse*, Lausanne : Delachaux & Niestlé.

Bronckard, Jean-Pierre (1996) : "Genre de textes, types de discours et opérations discursives", *Enjeux* 37-38, 31-47.

Dunlop, D. (1995) : "Practical considerations in the use of TEI headers in large corpora", *Computers and the Humanities* 29, 85-98.

Fleury, Serge / Folch, Helka / Habert, Benoît / Heiden, Serge / Illouz Gabriel / Lafon, Pierre / Prévost Sophie (2000) : "Profilage de textes : cadre de travail et expérience", *Actes du colloque 'JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles', Lausanne"*, 163-170.

Heiden, Serge (1999a) : *Weblex : manuel utilisateur*, UMR 8503, CNRS/ENS-LSH Lyon, `http://lexico.ens-lsh.fr/doc/weblex/`

Heiden, Serge (1999b) : "Encodage uniforme et normalisé de corpus : application à l'étude d'un débat parlementaire", *Mots* 60, 113-132.

Lafon, Pierre (1984) : *Dépouillements et statistiques en lexicométrie*, Genève-Paris : Slatkine-Champion.

Marchello-Nizia, Christiane (1999) : *Le français en diachronie*, Paris : Ophrys.

Prévost, Sophie / Heiden, Serge / Dupuis, Fernande (2000) : "Catégorisation d'un corpus hétérogène de français médiéval", *Actes du colloque 'JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles', Lausanne*", 485-492.