CORPUS

Corpus

7 | 2008
Constitution et exploitation des corpus d'ancien et de moyen français

Un corpus pour l'analyse de la variation et du changement linguistique

France Martineau

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/1508
DOI : 10.4000/corpus.1508
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 10 novembre 2008
ISSN : 1638-9808

Référence électronique
France Martineau, « Un corpus pour l'analyse de la variation et du changement linguistique », *Corpus* [En ligne], 7 | 2008, mis en ligne le 13 novembre 2009, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/1508 ; DOI : https://doi.org/10.4000/corpus.1508

Un corpus pour l'analyse de la variation et du changement linguistique[1]

France MARTINEAU
Université d'Ottawa

Cet article offre une réflexion sur la constitution de corpus pour l'analyse de la variation et du changement morphosyntaxique. L'analyse du changement linguistique doit tenir compte de facteurs internes au système linguistique, mais aussi de facteurs externes comme les conditions sociales et dialectales. L'étude de l'évolution de phénomènes morphosyntaxiques est facilitée par un corpus structuré intégrant des documents couvrant plus d'un état de langue. Ces critères – profondeur sociale, dialectale et historique – correspondent, comme nous le verrons, à un idéal à atteindre mais sont difficiles à respecter, surtout pour les textes médiévaux.

Nous présentons d'abord le projet *Modéliser le changement : les voies du français* (MCVF) puis nous abordons la structure du corpus. Nous discutons ensuite de l'enrichissement et de l'interrogation du corpus MCVF, de l'uniformisation des textes dans des formats standardisés et des principes qui sous-tendent l'élaboration de protocoles d'annotation morphosyntaxique. Enfin, nous terminons par une

1 Cet article a reçu l'appui financier du Conseil de recherches en sciences humaines du Canada dans le cadre du projet *Modéliser le changement : les voies du français*, dirigé par France Martineau (Grands travaux de recherche concertée). Cet article représente un travail d'équipe et nous tenons à remercier les chercheurs de notre équipe qui ont travaillé plus étroitement à l'enrichissement du corpus MCVF : Constanta Rodica Diaconescu, Monique Dufresne, Fernande Dupuis, Paul Hirschbühler, Anthony Kroch, Serge Lusignan, Christiane Marchello-Nizia, Yves Charles Morin et Beatrice Santorini. Des remerciements particuliers sont adressés à Paul Hirschbühler, à Christiane Marchello-Nizia et à Yves Charles Morin pour leurs commentaires et suggestions, ainsi qu'aux évaluateurs de la revue *Corpus*.

réflexion plus large sur le processus même d'élaboration de corpus.

1. Modéliser le changement : les voies du français

Nous savons tous que les langues changent et que ce changement est motivé par des facteurs internes et externes. Le processus lui-même ne nous est toutefois pas accessible ; seule la variation linguistique nous permet d'en percevoir les manifestations de surface. En d'autres mots, le changement ne se saisit que lorsqu'il fait surface.

Les changements qui affectent les phénomènes morphosyntaxiques s'étalent souvent sur plusieurs siècles, si bien que même en sociolinguistique moderne, lorsque des corpus en temps réel, avec des écarts de 20 ou 40 ans, sont utilisés, il est difficile de mesurer l'ampleur d'un changement en cours et son impact sur le système morphosyntaxique. Au contraire, la linguistique historique, parce qu'elle couvre souvent plusieurs états d'une langue, permet de saisir les mécanismes à l'œuvre dans le changement.

Toutefois, si les mécanismes qui sous-tendent l'évolution d'une langue sont universels, l'environnement linguistique sur lequel opèrent ces mécanismes reflète, de façon importante, le contexte social dans lequel ces langues se développent.

En linguistique historique, le texte nous est donné sans le contexte de production spontanée de la langue, qui pourrait servir d'étalon pour mesurer l'apport de facteurs comme le genre, le dialecte, la classe sociale du scripteur et le contexte de production de l'œuvre. Au contraire, en synchronie, il est toujours possible de mesurer l'écart entre la langue de l'œuvre et la langue vernaculaire. C'est donc l'un des défis de la linguistique historique : pouvoir non seulement comprendre les mécanismes internes du changement morphosyntaxique, mais également pouvoir articuler ce changement au contexte sociohistorique.

L'objectif principal visé par le projet *Modéliser le changement : les voies du français* est de mettre en évidence la dynamique des facteurs internes et externes, dans l'évolution

morphosyntaxique du français. Le projet est subventionné par le Conseil de recherches en sciences humaines du Canada, dans le cadre des projets majeurs des Grands travaux de recherche concertée. Nous avons obtenu pour ce projet un financement de 2,5 millions de dollars, étalé sur 5 ans (2005-2009).

Le projet s'appuie sur une équipe interdisciplinaire et regroupe des chercheurs de plusieurs universités canadiennes, européennes et américaines, qui sont responsables de certains de ses axes. Il est dirigé par France Martineau, à l'Université d'Ottawa. Y sont associés onze co-chercheurs, répartis dans sept universités, une quarantaine de collaborateurs, répartis dans plus de vingt universités, et des partenaires, en particulier les grands centres d'archives canadiens et des centres de recherche en linguistique informatique (voir le site Web : www.voies.uottawa.ca).

Afin de développer un modèle statistique et théorique du changement, nous avons établi dans ce projet un corpus annoté morphosyntaxiquement pour le français, qui couvre la période médiévale jusqu'au français classique, au moment où la structure morphosyntaxique du français moderne commence à se fixer.

Notre projet s'articule avec quatre autres projets majeurs de nature diachronique qui partagent avec notre projet les approches informatiques et théorique : *The Penn-Helsinki Parsed Corpus of Middle English* et *The Penn-Helsinki Parsed Corpus of Early Modern English* ; *The Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English* ; *Tycho Brahe Parsed Corpus of Historical Portuguese* ; et le *Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN)*.

Chacun de ces projets a pour objectif principal l'annotation morphosyntaxique de larges corpus, sur l'ancien anglais, le moyen anglais ou le portugais ancien. Bien que les décisions théoriques d'encodage soient indépendantes d'un projet à l'autre, nous cherchons à minimiser les écarts, de façon à pouvoir utiliser une même plate-forme d'interrogation et de comparaison de corpus de langues anciennes, de façon à mettre en évidence des propriétés fondamentales communes à toutes les langues et des propriétés variables qui particularisent l'histoire d'une langue sur une certaine période.

2. Corpus MCVF

Les documents numériques facilement accessibles sur le Web se sont multipliés lors des dernières décennies, laissant l'utilisateur avec le sentiment que la langue ancienne était là, au bout des doigts, prête à être analysée, qu'il ne suffisait plus que de télécharger les textes pour comprendre l'ancien ou le moyen français. Pourtant, cette masse de données doit pouvoir être structurée selon une logique propre à la linguistique historique afin de pouvoir révéler des régularités significatives sur le plan du changement linguistique. Seul un corpus d'envergure, structuré et étiqueté en fonction de paramètres linguistiques et sociaux, peut fournir des réponses à la dynamique qui a présidé à la formation de notre français actuel[2].

Le corpus développé dans le projet, plus de 2,5 millions de mots, se veut représentatif de la complexité des échanges dans la société française, du Moyen Âge jusqu'au XVIII[e] siècle, au moment où la morphosyntaxe connaît une forte pression de standardisation. Cette longue période du français que nous analysons – près de dix siècles d'histoire du français – permet de couvrir les grands phénomènes morphosyntaxiques qui ont affecté le français : le passage d'une langue V2 à une langue SVO ; la perte des sujets nuls ; la perte de la montée des clitiques objets ; la fixation du syntagme négatif et de la position des adverbes négatifs ; etc. Certains de ces phénomènes montrent peu de variation en ancien français, comme la montée du clitique objet, obligatoire pour tous les verbes modaux (Pearce 1990). Le changement s'amorce réellement en moyen français et prend de l'importance aux XVI[e] et XVII[e] siècles. Pour le suivre et voir sa diffusion d'une classe de verbe à l'autre, un corpus historique couvrant une période de temps suffisamment longue est nécessaire. Plus

2 Citons parmi d'autres, des projets pionniers dans la constitution de corpus pour l'analyse diachronique : le *Laboratoire de français ancien* de l'Université d'Ottawa, dirigé par Pierre Kunstmann et France Martineau, et surtout la *Base de français médiéval*, dirigée par Christiane Marchello-Nizia puis par Céline Guillot. Sur ce sujet, lire entre autres : Heiden et Lavrentiev (2004) ; Kunstmann (2000) ; Kunstmann, Martineau et Forget (2003) ; Marchello-Nizia (1999).

encore, ce changement a souvent été considéré comme lié à une série de phénomènes affectant les infinitives sur de longues périodes du français. La mise en évidence de changements affectant différentes structures, mais dont la diffusion aux dépens de l'ancienne structure se fait au même rythme dans le temps, est un indice du fait qu'à un niveau plus abstrait d'analyse grammaticale, ces divers changements reflètent un changement unique. Seul un corpus couvrant plusieurs états de langue peut dégager ces faisceaux de changement.

Comme l'un des objectifs du projet est également de comprendre l'impact des facteurs sociohistoriques, nous avons sélectionné les textes en fonction de leur représentativité dialectale ou régionale. Le corpus est divisé en quatre grandes zones dialectales : anglo-normand ; picard ; champenois / parisien ; centre-ouest. Pour les périodes précédant le français des XVII^{e} et XVIII^{e} siècles, la représentation dialectale est souvent tributaire de la disponibilité des textes et certaines régions sont nettement mieux représentées.

Nous avons inclus, dans notre analyse du changement linguistique, l'axe France – Nouvelle-France pour les XVII^{e} et XVIII^{e} siècles. L'une des préoccupations du projet est d'arriver à mieux expliquer comment s'est formé le français du Canada en partant de multiples sources régionales au moment de la colonisation aux XVII^{e} et XVIII^{e} siècles (voir Morin 2002 pour une hypothèse sur cette formation) et, de façon générale, de comprendre la dynamique des langues migrantes. Ces dernières fournissent des renseignements précieux sur l'impact des conditions sociohistoriques puisqu'elles présentent des configurations différentes des langues sources. Dans une perspective où l'histoire du français est aussi celle des français non hexagonaux, nous considérons que l'histoire du français au Canada est continue, depuis le Moyen Âge jusqu'à son implantation et son développement en Nouvelle-France.

Les corpus historiques, surtout s'ils couvrent plus d'une période historique, présentent une forte hétérogénéité. Les documents sont très courts pour la première période du français alors que les documents plus tardifs sont plus longs ; la variation dialectale de l'ancien français tend à s'amenuiser avec les siècles et les distinctions sont de nature plus régionale que

dialectale ; la variation sociale est peu représentée pour la période médiévale. À cette hétérogénéité inhérente à la disponibilité même des textes s'ajoute le fait que, dans le corpus MCVF, les documents proviennent de différents types d'édition de texte (éditions diplomatiques ou semi-diplomatiques) ou de manuscrits. Il serait envisageable de réduire cette hétérogénéité en limitant les critères de sélection : que des manuscrits ; que des textes littéraires ; que des textes picards ; etc. Toutefois, la nécessité d'avoir une masse critique de documents pour développer un modèle théorique du changement linguistique en s'appuyant sur des profils statistiques nous a incités à préférer une certaine hétérogénéité des textes.

3. Enrichissement et interrogation du corpus MCVF

3.1 Balisage du texte TEI

Nos documents proviennent de textes publiés, surtout pour la période médiévale et la Renaissance, ou plus rarement, de manuscrits. La transcription dans les deux cas obéit à un principe général de fidélité au texte d'appui. Dans le cas du document publié, nous avons sélectionné des éditions critiques de qualité, en omettant l'apparat critique. Dans le cas du manuscrit, nous avons reproduit sa disposition et l'orthographe. Le choix de s'en tenir le plus fidèlement possible à l'image du manuscrit permet d'examiner la question de l'émergence de normes régionales et parfois de retracer des prononciations.

Tous les textes du corpus MCVF ont d'abord été soumis à un balisage TEI (*Text Encoding Initiative*) de base, de façon à uniformiser les formats. Notre protocole de codage a pour source essentielle les *Principes directeurs pour l'encodage et l'échange de textes électroniques* (2007) élaboré par le consortium de TEI[3]. Comme source secondaire, nous avons

3 Burnard Lou & Syd Bauman. (2007), *Guidelines for Electronic Text Encoding and Interchange*, Version P5 ; <http://www.tei-c.org/Guidelines/P5/>

utilisé le manuel d'encodage de la Base de Français Médiéval élaboré par Heiden, Guillot et Lavrentiev[4].

Une partie des documents médiévaux provient d'échanges avec d'autres groupes de recherche, dont la *Base de français médiéval* (BFM). Le choix de formats standardisés TEI a permis d'accélérer le processus d'uniformisation des textes puisque nous avions choisi le même standard. Il faut toutefois reconnaître que, même à l'intérieur du protocole TEI, plusieurs choix s'offrent à l'utilisateur et que les choix peuvent différer de façon importante d'un projet à l'autre (voir notre manuel sur le site www.voies.uottawa.ca)[6].

Nous avons choisi des caractères Unicode pour rendre compte de certains signes, comme le ſ long. De même, nous avons utilisé des signes typographiques spécifiques pour signaler les cas d'agglutination dans l'original qui permettront aux outils de recherche de faire abstraction de cette agglutination (semblable à l'usage des apostrophes dans les éditions modernes de textes écrits ou imprimés avant le XVI^{e} siècle où *l'amour* indique que la version originale s'écrivait *lamour*, tout en faisant savoir au lecteur – et aux outils de recherche – que le *l* est une unité lexicale distincte de la suite graphique *amour* qui suit). De façon générale, nous avons fait le choix de ne pas utiliser des balises TEI à l'intérieur du mot, de façon à permettre aux logiciels de recherche et d'analyse linguistique de reconnaître plus facilement les suites de lettres qui constituent une unité lexicale. Le mot *lamour* est dès lors transcrit *l|amour*, au lieu d'être transcrit avec une balise indiquant la séparation des deux mots.

---

4 Heiden S., Guillot C. & Lavrentiev A. (2005) *Manuel d'encodage BFM / XML-TEI*, Version 2.1, BFM - Base de Français Médiéval [En ligne] ; Lyon : UMR ICAR / ENS-LSH <http://bfm.ens-lsh.fr/IMG/pdf/Manuel_Encodage_TEI.pdf>

5 La question de l'encodage TEI pour des corpus en linguistique historique a fait l'objet de nombreuses discussions. Citons Heiden & Lavrentiev (2004) ; Heiden & Guillot (2003).

6 La question de l'encodage TEI pour des corpus en linguistique historique a fait l'objet de nombreuses discussions. Citons Heiden & Lavrentiev (2004) ; Heiden & Guillot (2003).

Les documents, dans la version balisée, peuvent être consultés sur le site du projet (www.voies.uottawa.ca); l'interrogation est rendue possible grâce au logiciel *Philologic*, conçu à l'université de Chicago par l'équipe de ARTFL (*American Research on the Treasury of French Language*) dont nous avons adapté l'interface pour nos besoins.

Nous avons inclus, dans une base de données liée à l'en-tête TEI, des métadonnées permettant de documenter le profil de l'auteur, le parcours du texte (scribe ou imprimeur, région du document, par exemple) et les différents stades de traitement du texte avec leurs intervenants dans le projet. Cet enrichissement est essentiel pour articuler notre volet linguistique à notre volet historique. Toutes les informations de l'en-tête TEI sont intégrées dans une base de données avec les sources des informations recueillies (quel chercheur, dans quel article, a conclu qu'un tel document était représentatif de la langue d'une région donnée ?). L'intégration en-tête TEI / base de données / interface est en cours.

L'adaptation de l'interface de consultation nous a amenés à nous interroger sur les descripteurs et sur la stabilité de la notion qu'ils couvraient, selon les périodes visées. Ainsi, certains de ces descripteurs vont de soi – nom de l'auteur, lieu et date de naissance – mais leur importance dépend de la période visée.

L'importance de connaître le nom du scribe d'un texte, par exemple le scribe Guiot, est sans doute aussi grande que celle de connaître l'auteur Chrétien de Troyes pour l'étude d'un texte médiéval comme *Yvain ou le Chevalier au Lion* alors que cette question devient moins cruciale pour les textes classiques. De même, le lieu de naissance du scribe mais également le lieu où a été rédigé le document doivent être considérés au même titre que le lieu d'origine de l'auteur.

Pour la période médiévale, nous avons donc ajouté des informations sur le scribe, s'il était connu (nom, date et lieu de naissance), mais aussi sur le document (lieu et date de rédaction). Pour la Renaissance et la période classique, nous avons ajouté des possibilités de requête sur l'imprimeur (nom, date et lieu de naissance) ainsi que sur la maison d'impression (nom, date et lieu d'impression). Nous avons également ajouté

une possibilité de requête sur le métier de l'auteur ; si la question du « métier » de nombreux auteurs de l'époque médiévale ou de la Renaissance n'a guère d'effet sur la variation linguistique, cet aspect prend de l'importance pour la période classique, pour laquelle nous avons plus de textes présentant une stratification sociale et une variété de textes non littéraires.

Enfin, nous avons intégré une possibilité de requête par genre (vers / prose ; domaine littéraire / non littéraire) et par type de texte (chronique, roman, essai, etc.). Cette dernière catégorie a été particulièrement difficile à formaliser étant donné les genres hybrides ; ainsi, une relation de voyage relève-t-elle du domaine littéraire ou non littéraire ?

3.2 Annotation morphosyntaxique du texte

L'annotation morphosyntaxique du corpus se fait en deux étapes, étroitement liées l'une à l'autre : d'abord l'étiquetage des parties du discours puis celui des structures syntaxiques. Pour les étapes de l'annotation et les logiciels utilisés, nous renvoyons le lecteur à Martineau, Diaconescu & Hirschbühler (2007).

L'annotation morphologique, telle qu'elle est conçue dans le projet *Modéliser le changement : les voies du français*, est une étape vers l'annotation syntaxique, les deux étapes étant étroitement liées. Ce choix d'annoter morphologiquement en vue d'une annotation syntaxique explique certaines dispositions que nous avons prises.

Notre protocole d'annotation morphologique ne requiert qu'une annotation minimale des parties du discours, suffisante pour permettre une annotation syntaxique. Notre jeu d'étiquettes emploie généralement de 1 à 3 champs. Ainsi, le nom *filles* sera étiqueté NCPL, N pour la catégorie NOM, C pour le type COMMUN, et PL pour le pluriel, ce troisième champ étant généralement utilisé pour des phénomènes d'accord. Dans un souci de réduire l'information morphologique redondante, nous avons pris la décision d'indiquer le nombre sur le nom, *Les/D petites/ADJ filles/NCPL*, sachant que le nom, comme noyau du groupe nominal, pourra transmettre, lors de

l'annotation syntaxique, cette étiquette au groupe syntaxique NP.

Afin d'améliorer la performance de l'étiqueteur, nous avons séparé explicitement la préposition du pronom relatif dans les formes du type *auquel* ; dans un projet comme celui de la BFM, le protocole Cattext, orienté vers une annotation morphologique riche, sans annotation syntaxique, a choisi de créer une étiquette pour rendre compte de la forme morphologique agglutinée. C'est également une perspective morphologique et fonctionnelle qui a conduit à distinguer, dans le protocole Cattext, les pronoms relatifs des pronoms interrogatifs et des pronoms exclamatifs ; dans le protocole de notre projet *Voies du français*, les trois types de pronoms sont réunis sous une seule étiquette WPRO.

De même, certaines étiquettes morphologiques sont introduites afin de faciliter les distinctions syntaxiques. Comme Martineau, Diaconescu & Hirschbühler (2007) le soulignent, l'étiquetage de la préposition *à*, sur le plan morphologique, ne pose pas de problème. Toutefois, dans l'optique d'un étiquetage morphologique comme étape vers une annotation syntaxique, il nous a paru important de distinguer la préposition *à* selon qu'elle introduit un syntagme prépositionnel proprement dit ou un complément d'objet indirect. Nous avons donc décidé d'étiqueter *à* comme P dans le cas d'un syntagme prépositionnel proprement dit PP (1a), par opposition aux syntagmes gouvernés par la préposition *à* qui sont typiquement pronominalisables par les clitiques *lui* / *leur* (1b).

(1a) n'/NEG ay/AJ esté/VPP à/P la/D grande/ADJ feste/NCS

(1b) dist/VJ a/DAT son/DZ père/NCS

Cet étiquetage permet au logiciel de faire la distinction entre les deux syntagmes, le syntagme étiqueté comme PP et celui qui va être étiqueté comme NP-DTV. Une grande part de ces régularités sont lexicales, par exemple *penser à* est toujours suivi d'un PP. Toutefois, en l'absence d'un travail systématique sur cette question, nous avons choisi de traiter ces distinctions manuellement. Nous nous sommes heurtés à un problème semblable lors de l'étiquetage des pronoms accusatif et datif de certains verbes qui présentent une variation valencielle. Ainsi,

le verbe *prier*[7] en ancien français peut accepter un argument accusatif comme dans *Je le prie de venir* ou un argument datif dans *Je lui prie de venir*. Dans un cas ambigu comme *Il me prie de venir*, il est impossible de déterminer le cas du pronom *me* et nous avons opté pour l'accusatif.

La procédure de vérification manuelle est également conçue en fonction d'un étiquetage syntaxique. Tous nos documents subissent deux vérifications manuelles de l'étiquetage morphologique semi-automatique. La première vérification est faite sur la sortie du logiciel d'étiquetage morphologique semi-automatique ; toutefois, la deuxième vérification est faite en même temps que la première vérification de l'annotation syntaxique sur la sortie du logiciel d'étiquetage syntaxique semi-automatique. Cette procédure permet d'harmoniser les étiquettes morphologiques et syntaxiques.

Enfin, ce choix d'une annotation morphologique en vue d'une annotation syntaxique a des effets sur la définition des catégories morphologiques, sachant que le corpus couvre plusieurs états de langue, avec des changements de catégories morphologiques. Notre protocole d'étiquetage est stable, quelle que soit la période historique examinée.

Dans une perspective différente, où l'objectif principal est l'annotation morphologique ou une représentation du changement sémantique, il serait sans doute préférable de prévoir un jeu d'étiquettes différent, selon la période historique. Ainsi, certains adverbes de négation comme *rien* ont pu avoir en ancien français un sens positif (2a), puis être utilisés comme termes de polarité négative (2b, 2c) avant d'être interprétés comme des éléments négatifs (2c) (Buridant 2000, Martin 1996, Fournier 1998, Martineau & Déprez 2004). Tous ces éléments, sauf (2a) où *rien* est précédé d'un déterminant, sont toutefois étiquetés comme des adverbes de négation.

(2a) Et si vos dirai une *rien*
Si vuel que vos le saciés bien (Béroul, 179 ; cité par Moignet 1984 : 179)

7 Nous adoptons par commodité la graphie moderne.

(2b) Que voudriez-vous que je fisse? Je *n*'ay *pas rien* à faire (Bonaventure des Périers, *Nouvelle* 74 ; t. II, 45 ; cité par Martineau & Déprez 2004)

(2c) Est-ce que j'en sçay *rien* ? (Brécourt, *La Nopce de village* ; cité par Martineau & Déprez 2004)

(2d) Je ne veux *rien*.

L'étape qui suit l'annotation morphologique assigne des fonctions syntaxiques aux constituants, dans une représentation arborescente. Comme les étiquettes syntaxiques sont assignées à des nœuds dans l'arbre et non à des mots, ces nœuds portent l'étiquette syntaxique. Par rapport à une analyse morphologique de base, ce type d'étiquetage syntaxique offre plus d'un avantage : il permet d'identifier les fonctions, cela va de soi, mais également, dans le cadre théorique que nous adoptons, de représenter, par des indices dans une structure hiérarchique, des traces d'un élément qui a été déplacé ou encore d'indiquer des éléments occupant une position structurale, mais dont la réalisation phonétique est nulle. C'est le cas des sujets nuls, comme on peut le voir en (3).

(3) A mon advis, n'avions point esté plus de trois jours devant Paris quant le roy y entra. (Commynes)

```
(0 (1 IPMAT (2 PP (3 P A)
        (5 NP (6 DZ mon) (8 NCS advis)))
        (10 PON ,)
        (12 NPSBJ *pro*)
        (14 NEG n')
        (16 AJ avions)
        (18 ADVP (19 ADVNEG point))
                (21 VPP esté)
        (23 NPMSR (24 QR plus)
        (26 NP (27 DF de) (29 ADJNUM trois) (31 NCPL jours)))
        (33 PP (34 P devant)
                (36 NP (37 NPRS Paris)))
```

Ce type d'annotation syntaxique permet des recherches rapides de structures courantes en ancien et moyen français : position du verbe (V2/SVO), montée du clitique objet, expression du sujet. Le corpus annoté morphosyntaxiquement peut être

interrogé sur notre site Web grâce à Corpus Search dont nous avons adapté l'interface à nos besoins (www.voies.uottawa.ca/corpus_pg_fr.html).

Dans Martineau, Diaconescu & Hirschbühler (2007), nous avons illustré l'intérêt de Corpus Search pour l'analyse des sujets nuls à partir des livres 1 et 2 des *Mémoires* de Commynes. Nous avons montré que le pourcentage de sujets nuls est de 19 % (990/5219), que le contexte en proposition principale est généralement plus favorable à l'emploi des sujets nuls que le contexte en subordonnée, et que les sujets nuls ne sont pas restreints à une classe particulière de verbes comme les verbes impersonnels. Des recherches sur des mouvements syntaxiques, comme la montée du clitique objet, peuvent également aider à rapidement situer la fréquence du phénomène dans un texte, mais aussi selon les classes de verbes (verbes modaux comme *pouvoir* et *devoir* ou verbes aspectuels comme *commencer*).

Corpus Search permet aussi de cibler la position d'un élément par rapport à un autre, comme dans le cas de la position préverbale ou postverbale du sujet exprimé par rapport au verbe à temps conjugué ou de la position préverbale ou postverbale de l'adverbe de négation dans les infinitives.

Prenons le cas des adverbes de négation *pas* et *point*. On sait qu'au moins jusqu'en français classique, les deux adverbes se font concurrence[8], et que la variation est contrainte par des facteurs syntaxiques et régionaux (voir entre autres, Price 1997, Marchello-Nizia 1997, Martineau 2006, Volker 2007 et Kawaguchi 2008, sous presse). Pour mesurer la progression de *pas* et la fixation de sa position dans le syntagme négatif, on peut bien sûr se fonder sur les statistiques de fréquence de l'un et l'autre adverbe dans les textes. Ce type de résultats est facile à trouver dans des corpus annotés morphologiquement, où il suffit de faire une requête sur les adverbes de négation *pas* ou *point* pour trouver la fréquence relative d'un adverbe par rapport à l'autre, et la progression de

8 *Mie* fait également partie de cette alternance, au moins en ancien et en moyen français, avec probablement un emploi plus fréquent dans certaines régions du Nord et de l'Est. (cf. Völker 2003 ; 2007).

pas par rapport à *point* au fil des siècles. Il faudra néanmoins distinguer le contexte partitif du contexte non partitif, l'emploi de *point* demeurant privilégié dans le contexte partitif jusqu'en français classique.

(4a) Contexte non partitif :
ge ne cuit *pas* que gel connoisse. (*La Mort le Roi Artu*, 20.35 ; cité par Price 1997 : 174)

(4b) Contexte partitif :
si n'avoient *point* de viande. (Villehardouin 480 ; cité par Price 1997 : 174)

Mis à part la fréquence plus élevée de *pas*, à mesure que l'on se rapproche du français classique, la position préverbale de *pas* par rapport au verbe infinitif peut être considérée comme un autre signe de la fixation de l'adverbe dans le syntagme négatif. Comme l'ont montré Hirschbühler & Labelle (1994) et Martineau (1994), entre autres, l'adverbe de négation *pas* pouvait apparaître en position postverbale (5a, 5b). Peu à peu, la position préverbale s'est imposée, liant encore plus étroitement *ne…pas* dans le syntagme négatif (5c).

(5a) de ne combatre *pas* (Jouvencel, I, 184, cité par Martin & Wilmet 1980 : 27)

(5b) à ne le chercher *pas* (Navarre, *Heptaméron*, XVI^{e} s.)

(5c) de ne le *pas* aimer (La Fayette, *La Princesse de Clèves*, XVII^{e} s.)

On peut toutefois se demander si l'adverbe de négation *pas* s'est fixé plus rapidement en position préverbale dans le syntagme négatif que l'adverbe de négation *point*, étant donné que les deux adverbes sont en concurrence et que *point* a maintenu, plus tardivement, une valeur de terme de polarité.

Une recherche préliminaire sur la position de *point* dans les infinitives dans cinq textes déjà annotés morpho-syntaxiquement (*Yvain* de Chrétien de Troyes, *Mémoires* de Commynes, *Les Quinze joyes de mariage*, *Les Cent nouvelles nouvelles* ainsi que les *Lettres* de Marguerite de Valois) montre que les quelques occurrences de *point* dans les infinitives

apparaissent toutes en position postverbale (voir 6 et 7)[9]. C'est la norme en ancien et en moyen français comme en (6) dans *Les Quinze joyes de mariage*. On aurait toutefois pu s'attendre à une position préverbale dans les *Lettres*, datant du XVI[e] siècle, puisque déjà, à cette époque, *pas* commence à apparaître en position préverbale de l'infinitif.

(6) elle amast mieulx n'y aller point

```
NODE (IP-SUB (NP-SBJ (PRO elle))
                    (VJ amast)
                    (ADVP (ADVR mieulx))
                    (IP-INF (NEG n')
                        (PP (PRO y))
                        (VX aller)
                        (ADVP (ADVNEG point))))))
      (PONFP EOS))
  (ID XV-JOIES, 15.308))
```

(7) qui me prie ne m'annuier point de ses longeurs

```
NODE (CP-REL (WNP-1 (WPRO qui))
                    (IP-SUB (NP-SBJ *T*-1)
                        (NP-ACC (PRO me))
                        (VJ prie)
                        (IP-INF (IP-INF (NEG ne)
                                    (NP-RFL (PRO m|))
                                    (VX annuier)
                                    (ADVP (ADVNEG point))
                                    (PP (P de)
                                        (NP (DZ ses)
                                            (NCPL longeurs))))
  (ID VALOIS-AUTOGRAPH, 266.826))
```

Au XVI[e] siècle, la position postverbale demeure toutefois la plus fréquente, selon Gougenheim (1984) et Haase (1969) et comme le confirme l'analyse statistique de Hirschbühler et Labelle (1994). Ce n'est qu'au XVII[e] siècle, que la position

9 Il s'agit de montrer, à titre illustratif, le type de recherche qui peut être effectué avec Corpus Search. Il faudrait plus d'occurrences et de textes pour vérifier l'ampleur de ce patron. De plus, il faudrait analyser les contraintes syntaxiques (partitif / non partitif) et le profil individuel des textes en fonction de la région et de la période visées. Ce travail en cours n'est pas présenté dans cet article.

préverbale pour l'adverbe de négation commence à être plus fréquente; contrairement à ce que l'on aurait pu attendre, la position préverbale est alors privilégiée par *point* dans 41% des cas alors que *pas* se maintient bien en position postverbale, n'apparaissant que dans 15% des cas en position préverbale, selon les résultats de Hirschbühler & Labelle (1994). L'hypothèse d'une diffusion pré- ou post-verbale selon le statut de *pas* ou de *point* ne semble donc pas tenir; toutefois, il reste à vérifier si des contraintes sur les contextes partitif et non partitif ont pu avoir un effet sur la fréquence respective de *pas* et *point* dans les contextes préverbaux et postverbaux.

Ce type de recherche, sur la fixation de *pas* dans le syntagme négatif, montre la nécessité d'avoir des textes couvrant de longues périodes du français, en particulier des textes des XVII^{e} et XVIII^{e} siècles, au moment où se fixe la position préverbale des adverbes de négation. On peut également mesurer l'intérêt d'un corpus annoté morpho-syntaxiquement pour des recherches pointues où il faut non seulement comparer la fréquence d'une variante (*pas vs point*), mais aussi la fréquence dans certaines positions syntaxiques (propositions à temps conjugué *vs* infinitives ; position préverbale *vs* postverbale de l'infinitif).

Le modèle de changement linguistique que nous élaborons permet de dégager statistiquement les régularités de la langue et les écarts par rapport au système, et les changements qui sont en corrélation avec une même source, en particulier pour des changements en apparence indépendants (voir Kroch 1989). C'est ainsi qu'à mesure que les textes seront annotés, nous espérons pouvoir mesurer le lien entre la perte de V2 et celle des sujets nuls (voir Dufresne, à paraître, qui réunit des articles sur cette question, à partir d'analyses fondées sur le corpus MCVF).

4. Perspective de recherche

Le corpus MCVF est d'abord conçu pour l'analyse morphosyntaxique. Cet objectif détermine le choix de documents : documents longs, et donc le plus souvent littéraires, plutôt que brefs ; balisage TEI indiquant

essentiellement la mise en pages de base (page et ligne) plutôt qu'un balisage diplomatique ; et l'annotation morphologique de base.

En même temps, chaque étape de préparation du texte peut devenir une fin en soi. Ainsi, la préparation des textes, par exemple la résolution des agglutinations dans des documents à la graphie non standard, a facilité l'annotation morphologique de même que les recherches dans *Philologic*. Ce même travail sur les graphies non standard a été la source de recherches sur les stratégies d'écriture de scripteurs moins lettrés, comme certains scripteurs de l'exposition virtuelle *Les Canadas vus par les Canadiens 1750-1860*.

De même, la mise en place d'une base de métadonnées sur les descripteurs du texte est liée à l'enrichissement de l'en-tête TEI et à l'adaptation de l'interface de requêtes de *Philologic* ; cette même base de données est au cœur d'autres corpus (voir Martineau 2006 et Martineau, à paraître), ce qui permettra une standardisation des protocoles de saisie et d'interrogation.

L'annotation morphologique est à la source de l'annotation syntaxique ; en même temps, le corpus annoté morphologiquement ouvre la voie à la lemmatisation. Nous avons donc entrepris une collaboration avec l'ATILF, CNRS (Analyse et traitement informatique de la langue française) à l'Université de Nancy pour lemmatiser une partie du corpus médiéval et les corpus de peu-lettrés, en utilisant à la fois l'étiquetage morphologique du projet *Voies du français* et les ressources lexicales de l'ATILF. L'enrichissement des ressources lexicales devrait, indirectement, servir à la lemmatisation d'autres corpus.

Il semble de plus en plus nécessaire de privilégier des échanges et des collaborations entre les groupes de recherche qui travaillent dans la même direction. Des bases de travail uniformisées, comme l'est entre autres le balisage TEI, permet de faciliter les échanges. Sans cette standardisation, reprendre le travail du début est parfois plus rapide et efficace que d'investir temps et efforts dans la conversion de protocoles ou d'outils. Il semble donc important, dans l'élaboration de corpus destinés à la recherche de pointe, d'évaluer la compatibilité des outils et

des protocoles, l'adaptabilité à de nouvelles technologies et l'accessibilité des corpus à différents publics. C'est le paradoxe de la recherche de pointe : les outils développés doivent être adaptés à un objet de recherche précis tout en demeurant ouverts sur des développements futurs, difficiles à prévoir au moment de la réalisation du projet. L'importance de garder à l'esprit cette flexibilité des corpus est d'autant plus essentielle qu'elle assure la pérennité des corpus, dans les échanges ou lorsque les projets prennent fin et que les corpus sont assimilés à de plus grands ensembles.

Bibliographie

Sites Web

Analyse et traitement informatique de la langue française (ATILF), Nancy-Université, Centre national de la recherche scientifique (CNRS) et CMVF, *La base de textes lemmatisés* (http://stella.atilf.fr/gsouvay/mcvf/).

Daniel M. Bikel, *Software* (http://www.cis.upenn.edu/~dbikel/software.html).

Les Canadas vus par les Canadiens 1750-1860, pièces choisies de la Collection Baby (http://www.collectionbaby.uottawa.ca/).

Centro de Linguistica da Universidade de Lisboa, *Corpus Dialectal para o Estudo da Sintaxe* (CORDIAL-SIN) (http://www.clul.ul.pt/sectores/variacao/cordialsin/projecto_cordialsin.php).

fnTBL tagger (http://morphix-nlp.berlios.de/manual/node16.html).

Modéliser le changement : les voies du français (www.voies.uottawa.ca).

Penn Parsed Corpora of Historical English (http://www.ling.upenn.edu/hist-corpora/).

TEI : Text Encoding Initiative (http://www.tei-c.org/).

Tycho Brahe Parsed Corpus of Historical Portuguese (http://www.ime.usp.br/~tycho/corpus/files/index.html).

University of Chicago (The) – Laboratoire ATILF, *The ARTFL project* (http://humanities.uchicago.edu/orgs/ARTFL/).

York-Helsinki Parsed Corpus of Old English Poetry (The) (http://wwwusers.york.ac.uk/~lang18/pcorpus.html).

Buridant C. (2000). *Grammaire nouvelle de l'ancien français*. Paris : SEDES.

Dufresne M. (éd.) (à paraître). T*ypologie, ordre des mots et groupe verbal en français médiéval***.** Québec : Les Presses de l'Université Laval.

Fournier N. (1998). *Grammaire du français classique.* Paris : Belin.

Gougenheim G. (1984) [1951]. *Grammaire de la langue française du seizième siècle.* Paris : Picard.

Haase A. (1969). *Syntaxe française du XVII^e^ siècle*. Paris : Delagrave.

Heiden S. & Guillot C. (2003). « Capitalisation des savoirs par le Web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval ». In P. Kunstmann, F. Martineau & D. Forget (éds), *Ancien et moyen français sur le web. Enjeux méthodologiques et analyse du discours.* Ottawa : Les Editions David : 77-92.

Heiden S. & Lavrentiev A. (2004). « Ressources électroniques pour l'étude des textes médiévaux : approches et outils », *Revue Française de Linguistique Appliquée* IX, 1 : 99-118.

Hirschbühler P. & Labelle M. (1994). « Changes in Verb Position in French Negative Infinitival Clauses », *Language Variation and Change* 6 : 149-178.

Kawaguchi Y. (2008). « Particules négatives du français : ne, pas, point et mie : Un aperçu historique ». In L. Baronian & F. Martineau (eds), *Le français : d'un continent à l'autre*. Québec : Les Presses de l'Université Laval.

Kroch A. (1989). « Reflexes of Grammar in Patterns of Language Change », *Language Variation and Change* I, 3 : 199-244.

Kunstmann, P. (2000). « Ancien et moyen français sur le web : textes et bases de données », *Revue de Linguistique Romane* 64 : 17-42.

Kunstmann P., Martineau F. & Forget D. (éds) (2003). *Ancien et moyen français sur le web. Enjeux méthodologiques et analyse du discours.* Ottawa : Les Editions David.

Marchello-Nizia C. (1997[2]) [1979]. *Histoire de la langue française aux XIV^e^ et XV^e^ siècles*. Paris : Nathan.

Marchello-Nizia C. (1999). « Corpus diachroniques », *Revue française de linguistique appliquée* IV, 1 : 31-39.

Martin R. (1966). *Le mot « rien » et ses concurrents en français (du XIV^e^ siècle à l'époque contemporaine)*. Paris : Klincksieck.

Martin R. & Wilmet M. (1980). *Manuel du français du Moyen Age. 2. Syntaxe du moyen français*. Bordeaux : Sobodi.

Martineau F. (1994). « Movement of Negative Adverbs in French Infinitival Clauses », *Journal of French Language Studies* IV, 1 : 55-73.

Martineau F. (2006). « Perspectives sur le changement linguistique : aux sources du français canadien », *Revue canadienne de linguistique L*, 1-4 : 1001-1040. Numéro spécial du 50^e^ anniversaire de la revue.

Martineau F. (à paraître). « À distance de Paris : usages linguistiques en France et en Nouvelle-France à l'époque classique ». In D. Aquino-Weber, S. Cotelli & A. Kristol (eds), *Sociolinguistique historique du domaine gallo-roman. Enjeux et méthodologies d'un champ disciplinaire émergent.* Actes du colloque de Neuchâtel, Université de Neuchâtel, 8-9 juin 2007. Neuchâtel : Université de Neuchâtel, Centre de dialectologie et d'étude du français régional.

Martineau F. & Déprez V. (2004). « *Pas rien* / *Pas aucun* en français classique : variation dialectale et historique », *Langue française* 143 : 33-47. Numéro thématique « La négation en français classique ».

Martineau F., Diaconescu C. R. & Hirschbühler P. (2007). « Le Corpus Voies du français : de l'élaboration à l'annotation ». In P. Kunstmann et A. Stein (eds) *Le Nouveau Corpus d'Amsterdam*. Actes de l'atelier de Lauterbad, 23-26 février 2006. Stuttgart : Steiner, 121-142.

Moignet G. (1984[2]) [1973]. *Grammaire de l'ancien français. Morphologie, syntaxe*. Paris : Klincksieck.

Morin Y. C. (2002). « Les premiers immigrants et la prononciation du français au Québec », *Revue québécoise de linguistique* XXXI, 1 : 39-78.

Pearce E. (1990). *Parameters in Old French Syntax : Infinitival Complements*. Dordrecht : Kluwer Academic Publishers.

Price G. (1997). « Negative Particles in French ». In S. Gregory & D. A. Trotter (eds) *De mot en mot. Aspects of Medieval Linguistics*. Cardiff : University of Wales Press, 173-190.

Völker H. (2003). *Skripta und Variation. Untersuchungen zur Negation und zur Substantivflexion in altfranzösischen Urkunden der Grafschaft Luxemburg (1237–1281)*. Tübingen : Niemeyer.

Völker H. (2007). « A 'practice of the variant' and the origins of the standard. Presentation of a variationist linguistics method for a corpus of Old French charters », *French Language Studies* 17 : 207–223.