Presto, un corpus diachronique pour le français des XVI^{e}-XX^{e} siècles

Peter Blumenthal[1], Sascha Diwersy[2], Achille Falaise[3],
Marie-Hélène Lay[4], Gilles Souvay[5], Denis Vigier[6]
(1) Université de Cologne, Albertus-Magnus-Platz, D-50923 Köln, Allemagne
(2) PRAXILING, Univ. Montpellier-3, Route de Mende 34199 Montpellier, France
(3) ICAR, CNRS, 15 Parvis René Descartes, 69342 Lyon, France
(4) FoReLL, Univ. de Poitiers, 5 rue Théodore Lefebvre, 86073 Poitiers, France
(5) ATILF-CNRS, Univ. Nancy-2, 44 av. de la Libération, 54063 Nancy, France
(6) ICAR, Université de Lyon, 18 quai Claude Bernard, 69365 Lyon, France
peter.blumenthal@uni-koeln.de,sascha.diwersy@univ-montp3.fr,
achille.falaise@ens-lyon.fr, marie-helene.lay@univ-poitiers.fr,
gilles.souvay@atilf.fr, denis.vigier@univ-lyon2.fr

RÉSUMÉ

Le corpus Presto est un corpus diachronique du français couvrant la période XVI^{e}-XX^{e} siècles, annoté en étiquettes morphosyntaxiques, lemmes, et dépendances syntaxiques. Une partie de ce corpus (53 textes, 6,8 millions de mots) sera diffusée sous licence libre, ainsi que les ressources ayant permis cette annotation.

ABSTRACT

Presto, a Diachronic Corpus for the French of the XVIth-XXth Centuries

The Presto corpus is a French diachronic corpus covering the XVIth-XXth centuries, annotated with morphosyntactic tags, lemma, and syntactic dependencies. A part of this corpus (53 texts, 6.8 million words) will be distributed under a free license, as well as the resources allowing this annotation.

MOTS-CLÉS : corpus, annotations linguistiques, langue française, diachronie.

KEYWORDS: corpus, linguistic annotations, French langage, diachrony.

1 Introduction

Le français des XVI^{e}-XVIII^{e} siècles est un état de la langue assez proche du français moderne, du moins si on le compare au français médiéval. Durant cette période qui voit le français s'imposer comme une langue d'État normalisée, ce sont surtout les variantes graphiques qui posent problème pour l'analyse automatique (cf. figures 1 et 2). Ainsi, quoique les textes puissent apparaître relativement compréhensibles par un lecteur moderne non spécialiste, ils ne peuvent pas être traités automatiquement en exploitant les ressources existantes. Cet état de la langue est encore peu traité,

contrairement à la période médiévale, pour laquelle des ressources ont été créées ces dernières années. Ainsi, la *Base de Français Médiéval* (ENS de Lyon, Laboratoire ICAR, 2012) et le *Nouveau Corpus d'Amsterdam* (Gleßgen & Vachon, 2010) offrent des corpus étiquetés en parties du discours pour la période IX^{e}-XV^{e} siècles. À l'opposé, les corpus annotés du français moderne sont nombreux, mais la base *Frantext* catégorisée, par exemple, ne remonte pas avant 1850. Entre ces deux périodes, on peut citer principalement deux corpus partiellement étiquetés en parties du discours : quatre textes du corpus MCVF (« *Modéliser le changement : les voies du français*, corpus annoté XML », sous la direction de F. Martineau, avec P. Hirschbühler, S. Lusignan, C. Marchello-Nizia, Y. Ch. Morin et F. Rouget), et quatre textes des BVH (*Bibliothèques Virtuelles Humanistes* – L. Bertrand, M.-L. Demonet), ce qui est loin de représenter une couverture approfondie de cette période.

SI LA NATURE (dont quelque Person- de grand'renommée non sans rayson a douté, si on la devoit appeller Mere, ou Maratre) eust donné aux Hommes un commun vouloir, & consentement, outre les innumerables commoditez, qui en feussent procedées, l'Inconstance humaine, n'eust eu besoing de se forger tant de manieres de parler. Laquéle diversité, & confusion, se peut à bon droict appeller la Tour de Babel.

FIGURE 1: Début de *La deffence, et illustration de la langue francoyse*, Joachim du Bellay, 1549.

La langue françoise ne commença à prendre quelque forme que vers le dixieme siecle; elle naquit des ruines du latin & du celte, mêlées de quelques mots tudesques. Ce langage étoit d'abord le romanum rusticum, le romain rustique; & la langue tudesque fut la langue de la cour jusqu'au tems de Charles-le Chauve.

FIGURE 2: Extrait de l'article « FRANÇOIS, ou FRANÇAIS », rédigé par Voltaire, de l'*Encyclopédie*, 1751-1772.

Le projet franco-allemand Presto[1] (financé par l'ANR et la DFG pour la période 2013-2017) vise à mettre en œuvre des calculs statistiques automatisés pour étudier l'évolution sémantique et discursive, en diachronie, d'un ensemble de prépositions simples du français, en s'appuyant sur l'exploration quantitative de leurs cotextes distributionnels d'apparition.

À cette fin, nous avons construit un corpus étiqueté en parties du discours, lemmes et dépendances syntaxiques, visant à combler le manque de corpus étiquetés pour la période XVI^{e}-XVIII^{e} siècles, mais couvrant aussi la période XIX^{e}-XX^{e} siècles. L'annotation en parties du discours et en lemmes est effectuée à l'aide du logiciel *TreeTagger* (Schmid, 1994), entraîné sur un corpus annoté manuellement de 62k mots (80 % pour l'apprentissage, 10 % pour les tests, 10 % pour l'évaluation), et utilisant un lexique de formes anciennes, obtenues par archaïsation d'un lexique moderne. Une annotation en dépendances est en cours, basée sur la plateforme *Bonsai* (M. Candito, B. Crabbé, P. Denis, M. Falco, F. Guérin, E. Henestroza Anguiano, J. Nivre, D. Seddah – Candito & *al*. 2010).

2 Contributeurs

Le corpus Presto a été réalisé sous la direction de P. Blumenthal (Université de Cologne) et D. Vigier (Université de Lyon). Les textes qui y sont réunis ont été sélectionnés par P. Blumenthal, V. Goossens et D. Vigier, dans des bases textuelles existantes : Frantext (V. Montémont, G. Souvay),

[1] http://presto.ens-lyon.fr

les BVH (*Bibliothèques Virtuelles Humanistes* – L. Bertrand, M.-L. Demonet), l'ARTFL (American and *French Research on the Treasury of the French Language* – R. Morrissey, M. Olsen) et le CEPM (*Corpus électronique de la première modernité*).

Le jeu d'étiquettes et les conventions d'annotation ont été définis collégialement par les membres du projet Presto, à partir du jeu d'étiquettes MULTEXT/EAGLES (Leech & *al.* 1994) et GRACE (Adda & *al.* 1998).

Deux types de ressources informatiques ont été utilisés pour l'analyse automatique des textes : un lexique associant à chaque forme un lemme et une étiquette morphosyntaxique, et un corpus de référence (62k mots) analysé manuellement par des experts. Ces deux ressources servent à construire un modèle de langage pour l'analyse des textes. Le lexique se base principalement sur le *Lefff* (B. Sagot, 2010) et *Morphalou* (S. Salmon-Alt, L. Romary, J.-M. Pierrel), avec des formes archaïsées grâce au logiciel LGeRM (lemmatisation de la variation graphique des états anciens du français et lexiques morphologiques, G. Souvay). Le corpus de référence (sélection de cinq textes représentant des genres et des périodes différentes) a été pré-étiqueté par A. Falaise, avant d'être corrigé par M. Goux, C. Jacquot et D. Vigier, travaillant en parallèle sur les mêmes textes, à l'aide du logiciel Analog (M.-H. Lay). La fusion de ces annotations a été effectuée par A. Falaise (fusion automatique des cas triviaux) et D. Vigier (fusion experte pour les divergences d'annotation non triviaux). Ces ressources ont permis à S. Diwersy et A. Falaise de réaliser des modèles de langage pour l'annotation automatique du corpus.

3 Le corpus

3.1 Sélection des textes

Dans le corpus Presto, nous avons sélectionné des textes du XVI^{e} au XX^{e} siècles (jusqu'à 1944, pour des raisons de droit d'auteur) dans des bases textuelles existantes, en privilégiant les premières éditions des œuvres (notamment lorsqu'elles ont été publiées du vivant de l'auteur) ou, à défaut, celles respectant le plus l'orthographe d'époque. Cette sélection s'est en outre opérée sur le critère du genre textuel ; nous avons ainsi choisi de distinguer empiriquement quatre grands « genres » discursifs : narratif, poésie, théâtre et traité, et d'équilibrer le nombre de textes entre ces genres. Le corpus est structuré en trois ensembles concentriques :

- Un « **Corpus Presto noyau** » de 53 textes libres[2] : 34 textes des XVI^{e}-XVIII^{e} siècles et 19 textes des XIX^{e}-XX^{e}, jusqu'à 1944 (date au-delà de laquelle l'œuvre ou l'exemplaire peuvent tomber sous le coup des droits d'auteur ou d'éditeur). Ce corpus est équilibré sur le plan chronologique et générique, et totalise 6,8 millions de mots.

- Un « **Corpus Presto équilibré** » qui réunit 162 textes (les 53 textes libres du corpus noyau et 109 textes non libres), dont les dates d'édition s'échelonnent de 1509 (date d'édition de l'exemplaire) à 1944. Comme le corpus noyau, ce corpus est équilibré, mais il est de plus échantillonné : afin d'atténuer les biais statistiques dus aux différences de taille entre les textes, il ne comporte qu'un maximum de 50 000 mots par texte. Concrètement, chaque texte est découpé en 5 parties, et on ne retient pour chaque partie qu'une portion centrale

[2] Licence *Creative Commons 3.0 BY-SA-NC*.

de 10 000 mots environ (les phrases ne sont pas coupées). De ce fait, bien que comportant plus de textes que le corpus noyau, il ne comporte que 5,4 millions de mots.

- Un **« Corpus Presto étendu »**, non échantillonné, et moins équilibré sur les plans temporel et générique, puisqu'ont été agrégés selon les opportunités jusqu'à quatre textes supplémentaires par décennie.

Corpus	**Période**	**Genres**	**Taille**	**Licence**
Noyau	XVI^{e}-XX^{e} s.	narratif, poésie, théâtre, et traité	53 textes, 6,8 M mots	CC 3.0 BY-SA-NC
Équilibré			162 textes, 5,4 M mots	Sous droits
Étendu	XVI^{e}-XXI^{e} s.		340 textes, 35,5 M mots	Sous droits

TABLE 1 : Principales versions du corpus Presto.

Hors des corpus « Presto », nous constituons en outre deux corpus spécialisés destinés à des études longitudinales plus spécifiques portant sur un seul type de genre discursif : la presse française[3] du XIX^{e} et du XX^{e} siècle d'une part, le discours encyclopédique[4] (XVIII^{e}-XXI^{e} s.) d'autre part. La plupart des textes de ces corpus ne sont malheureusement pas libres, et ne peuvent donc pas être redistribués.

3.2 Type d'annotation

Ces textes sont annotés en morphosyntaxe, en lemmes, et en dépendances, de manière homogène du XVI^{e} au XX^{e} siècles, selon un méthode détaillée dans (Diwersy & *al.*, 2015 ; Diwersy & *al.*, 2017).

3.2.1 Annotations structurelles

Dans la mesure où le corpus Presto comporte une diversité notable de genres textuels, nous avons défini un modèle d'annotation structurale facilement exploitable, mais largement sous-spécifié. Ainsi, l'unité de base de ce corpus est le *texte* (comportant un identifiant, le titre, l'auteur, l'année de parution, le genre textuel et la thématique), lui-même seulement divisé en *phrases* typogrtaphiques, faisant en cela abstraction de toute unité de structuration intermédiaire, comme par exemple les chapitres ou paragraphes. À titre de comparaison, le corpus complémentaire des encyclopédies, homogène sur le plan générique, permet une structuration plus détaillée : le corpus est divisé en *entrées* (comportant l'ouvrage source, le lemme de l'article, l'année de parution et la classification thématique) ; chaque entrée est elle-même structurée en *tête de l'entrée* et *corps du texte* ; ce dernier se divise en paragraphes, eux-mêmes constitués de *phrases*.

---

[3] Presse nationale : *Le Figaro* (1826-1895, 2002), le *Journal des Débats* (1830-1892), *La Presse* (1840-1900), *Le Monde* (2002), *Sud-Ouest* (2002). Presse locale (Lyon) : *L'Éclair* (1881-1885), *Le Passe-Temps* (1879-1899), *La Renaissance* (1877-1884), *Le Réveil de Lyon* (1881-1882). Presse locale (Nord-Est de la France) : *L'Est Républicain* (2002).

[4] En cours de construction ; actuellement 942 articles communs à l'*Encyclopédie* (XVIII^{e} s.), l'*Encyclopædia Universalis* (XX^{e} s.) et *Wikipedia* (XXI^{e} s.).

3.2.2 Tokenisation

Le problème de l'identification de « mots », déjà délicat en français moderne, s'amplifie en diachronie ; on peut ainsi observer des formes anciennes qui agglutinent deux formes modernes (par exemple *trèsardent*), ou au contraire des formes anciennes qui ne se sont pas encore agglutinées (*par ce que*, *par cela que*, *Ménil-Montant*, *n'a guere*, *na gueres*). Il est possible de faire cohabiter au sein d'un corpus plusieurs niveaux de segmentation, mais les algorithmes classiques de traitement automatique des langues fonctionnent sur des *tokens* univoques. La tokenisation du corpus Presto cherche à atteindre trois objectifs :

1. La tokenisation doit être univoque, c'est à dire sans ambiguïté et avec un seul niveau de tokenisation.
2. La tokenisation doit être facilement automatisable, et donc reposer sur des informations de bas niveau, comme la graphie et le lexique (liste de formes existantes), et non sur des connaissances morphologiques ou étymologiques.
3. Enfin, la tokenisation obtenue doit être facilement utilisable pour un locuteur du français moderne non spécialiste des autres périodes.

Par défaut, nous appliquons une segmentation « maximale »[5], en nous basant sur la typographie (espaces, etc.), dès lors que chacun des tokens ainsi créé peut se voir attribuer une étiquette morphosyntaxique. Ainsi, nous distinguons un seul token pour *na guere* (forme de *naguère*, où *na* n'a pas d'étiquette), deux tokens pour *lors que*, et trois tokens pour *par ce que*. Dans la même logique, nous segmentons les unités typographiques qui sont séparées en français moderne, comme par exemple *tresgrand* ou *moymesme* (deux tokens). Par contre, *ledit*, qui n'est pas séparé en français moderne, reste pour nous un seul token. Cette approche pragmatique peut se révéler fautive : sans correction manuelle, nous risquons ainsi de segmenter la forme *tressaillant* en deux tokens, même quand il s'agit du verbe *tressaillir*. En outre, nous traitons la locution concessive *bien que* comme deux tokens.

3.2.3 Étiquettes morphosyntaxiques

Pour Presto, nous avons adapté les jeux d'étiquettes MULTEXT/EAGLES et GRACE, déjà largement utilisés dans d'autres projets, en les adaptant aux objectifs du projet, mais aussi dans la perspective de l'annotation à venir. En effet, le choix d'un jeu d'étiquettes a de multiples répercussions sur la campagne d'annotation et sur l'exploitabilité des résultats : plus le jeu d'étiquettes est précis, plus la phase de désambiguïsation est complexe ; d'une part, parce que la plus grande précision du jeu d'étiquettes génère des ambiguïtés, multipliant ainsi les choix à faire et à valider ; d'autre part parce qu'il augmente la nécessité d'un annotateur expert capable de faire des choix complexes.

En diachronie, les choses sont plus complexes encore, car l'organisation syntaxique évolue, ce qui conduit à des lectures différentes de séquences comparables en surface (point particulièrement sensible pour les participes[6]), et les unités peuvent se trouver recatégorisées dans le temps (des noms deviennent préposition, d'autres changent de genre, etc). Or, nous voulons disposer d'un jeu d'étiquettes pertinent pour toute la période, et qui permette une désambiguïsation « fiable » à

[5] Les règles détaillées sont disponibles sur le site du projet Presto.

[6] En français classique, la classification des formes en *–ant* ne va pas de soi dans la mesure où elle n'est pas marquée formellement par une morphologie distinctive : le gérondif, invariable, se distingue mal du participe (au masculin singulier) du fait qu'il n'est pas régulièrement précédé de *en* ; le participe qui peut être variable en genre et en nombre, se distingue mal de l'adjectif verbal (Fournier 2002, §421 : 291-292).

« faible coût cognitif ». Il nous fallait donc choisir un jeu d'étiquettes qui, s'il n'est pas très riche, permette de faire des choix qui ne grèvent pas de futures interprétations, qui pourront se manifester lors d'étiquetages ultérieurs plus fins si le besoin s'en fait sentir. Ces considérations nous ont conduits à opter pour un jeu d'étiquettes « à grain faible ».

Catégorie	**Définition**	**Sous-catégories (définitions)**	
N	Nom	c (commun), p (propre)	
V	Verbe	u (être/avoir), v (autre verbe)	c (conjugué), n (infinitif)
A	Adjectif	g (général), p (possessif)	
P	Pronom	p (personnel), d (démonstratif), i (indéfini), s (possessif), t (interrogatif), r (relatif)	
D	Déterminant	a (article défini), d (démonstratif), n (article indéfini), p (article partitif), i (indéfini), r (relatif), t (interrogatif/exclamatif)	
G	Participe-Adjectif-Gérondif	a (part. présent/adjectif verbal/gérondif), e (part. passé/adjectif verbal)	
R	Adverbe	g (général), p (particule), t (interro-exclamatif)	
S	Adposition	-	
C	Conjonction	c (coordination), s (subordination)	
M	Numéral	c (cardinal), o (ordinal)	
I	Interjection	-	
X	Résidu	a (abréviation), e (mot étranger), s (symbole), p (préfixe détaché), i (consonne intercalée)	
F	Ponctuation	s (forte), w (faible), o (autre)	

TABLE 2 : Jeu d'étiquettes morphosyntaxiques *Presto_min* du corpus Presto.

Ce jeu d'étiquettes « réduit » (*Presto_min*) reprend donc les catégories classiques, sans annotation flexionnelle, en faisant le choix de regrouper dans un premier temps les étiquettes dont la grande précision demande une désambiguïsation experte, ne serait-ce que parce qu'elle présente une forte variabilité en diachronie. On a ainsi choisi d'isoler une zone particulièrement délicate pour les périodes préclassique et classique essentiellement, en créant une étiquette qui se démarque de la typologie traditionnelle : l'étiquette *G* nous permet de regrouper les divers emplois des participes, sans avoir à nous prononcer sur la catégorie à laquelle ils appartiennent en contexte (forme verbale ou adjectivale ?). En effet, contraindre à faire ce choix (sujet à erreur et à discussion des interprétations sous-jacentes) pose le problème du rattachement au lemme, et nuit à la progression du travail d'annotation : les étiquetages « erronés » sont difficiles à localiser, et on ne peut garantir l'achèvement de la désambiguïsation des étiquettes verbes/adjectifs. Nous avons donc préféré isoler ce cas afin de pouvoir lui consacrer des traitements spécifiques. Ce choix a un impact sur le statut auxiliaire ou verbe plein des verbes *être/avoir* : le participe appelle l'auxiliaire, pas l'adjectif. Nous avons donc choisi de neutraliser cette distinction : nous n'avons pas d'étiquette spécifique pour les auxiliaires, qui sont codés comme des verbes.

3.2.4 Lemmatisation

Dans un lexique du français ancien, la variation graphique rend les cas d'ambiguïté lexicale (plusieurs entrées pour une même forme) beaucoup plus fréquents qu'en français moderne. Une lemmatisation du corpus s'avère donc particulièrement utile en vue de son exploitation linguistique. Nous utilisons des entrées modernes issues de deux lexiques morphologiques existants : le *Lefff* (Sagot, 2010) et *Morphalou* (Romary *& al.*, 2004). Toutefois, la langue des XVI^{e}-XVIII^{e} siècles utilise aussi un vocabulaire ancien qui n'a pas d'équivalent moderne. Cette base moderne a donc été complétée avec les lexiques morphologiques anciens liés au lemmatiseur *LGeRM* (Souvay & Pierrel, 2009), pour les lemmes sans équivalent moderne relevés sur les corpus Presto et Frantext. La base lexicale ainsi obtenue se veut exhaustive sur le plan des lemmes, mais ne leur associe quasiment que des formes modernes. Nous utilisons donc la base de règles du système LGeRM afin d'archaïser ces dernières.

3.2.5 Dépendances syntaxiques

L'analyse en dépendances du corpus est en cours de réalisation. Elle s'appuie sur la plateforme d'analyse en deux étapes *Bonsai*, et utilise l'analyse Presto à la place de l'analyse Bonsai pour la première étape (analyse morphosyntaxique), et l'analyse Bonsai pour la deuxième étape (dépendances syntaxiques). Les dépendances du corpus suivent donc les conventions d'étiquetage en dépendances du *French Treebank* (Abeillé & *al*., 2003), sur lequel la plateforme Bonsai est entraînée.

3.3 Distribution

Le corpus noyau, sa documentation, ainsi que les ressources élaborées pour son annotation, sont disponibles sous licence *Creative Commons 3.0 NC-BY-SA*, et seront progressivement diffusés sur le site du projet courant 2017.

4 Discussion

Les ressources développées dans le cadre de Presto permettent de construire un modèle pour TreeTagger, qui obtient un taux d'étiquetage morphosyntaxique correct (*précision*) de 94,6 % pour l'ensemble de la période XVI^{e}-XVIII^{e} siècles, mais qui s'avère sensiblement plus bas pour le XVI^{e} siècle (91,4 %). Pour ce siècle, le développement de ressources plus spécifiques permettrait probablement d'améliorer ce taux. Il faut aussi noter que la désambiguïsation ne porte que sur les étiquettes, et non sur les lemmes. Cela pose rarement problème en français moderne, où l'orthographe permet généralement de distinguer les lemmes, mais, dans notre cas, il nous est par exemple impossible de distinguer le lemme de « *congres* » (*CONGRE ou CONGRÈS*) ; l'application d'un algorithme de désambiguïsation comme (Lesk, 1986) permettrait de résoudre ce problème.

Remerciements

Ce travail est issu du projet Presto, cofinancé par l'*Agence Nationale de la Recherche* (ANR) et la *Deutsche Forschungsgemeinschaft* (DFG).

Références

ABEILLÉ, A., CLÉMENT L., TOUSSENEL F. (2003). Building a treebank for French, in A. Abeillé (ed) *Treebanks*, Kluwer, Dordrecht.

ADDA G., MARIANI J., LECOMTE J., PAROUBEK P., RAJMAN M. (1998). « The GRACE French Part-of Speech Tagging Evaluation Task ». Proceedings of the *First International Conference on Language Resources and Evaluation* (LREC 1998), p. 433-441.

CANDITO M.-H., NIVRE J., DENIS P., HENESTROZA ANGUIANO E. (2010). « Benchmarking of Statistical Dependency Parsers for French ». Proceedings of *COLING'2010*, Pékin, Chine.

DIWERSY S., FALAISE A., LAY M.-H., SOUVAY G. (2017). Ressources et méthodes pour l'analyse diachronique. *Langages* 206, à paraître.

DIWERSY S., FALAISE A., LAY M.-H., SOUVAY G. (2015). Traitements pour l'analyse du français préclassique. Actes de *Traitement Automatique des Langues Naturelles* (TALN 2015), Caen.

FOURNIER N. (2002). *Grammaire du français classique*. Paris : Belin sup.

GLEßGEN M.-D., VACHON C. (2010). *Répertoire bibliographique du Nouveau Corpus d'Amsterdam, établi par Anthonij Dees et Piet Van Reenen (Amsterdam 1987), revu et élargi par M.-D.G. et C.V.*, 3. ed. Stuttgart: Institut für Linguistik/Romanistik.

IDE N., VÉRONIS J. (1994). « MULTEXT: Multilingual Text Tools and Corpora ». Proceedings of the *15th International Conference on Computational Linguistics* (COLING'94), Kyoto, Japon.

LAY M.-H., PINCEMIN B. (2010). Pour une exploration humaniste des textes : AnaLog. Actes des *Journées internationales d'Analyse statistique des Données Textuelles*, Rome, Italie.

LEECH G., WILSON A. (1994). *EAGLES Morphosyntactic Annotation, Draft - Work in Progress*. Rapport technique, Lancaster, EAG-CSG/IR-T3.1.

LESK, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of the *5th annual international conference on Systems documentation* (SIGDOC '86), pages 24-26, New York, NY, États-Unis. ACM.

SAGOT B. (2010). « The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French ». Proceedings of the *7th international conference on Language Resources and Evaluation* (LREC 2010), Istanbul, Turquie.

SCHMID H. (1994). « Probabilistic Part-of-Speech Tagging Using Decision Trees ». Proceedings of *International Conference on New Methods in Language Processing*, Manchester, Royaume-Uni.

SOUVAY G., PIERREL J.-M. (2009). LGeRM : lemmatisation de mots en moyen français. *Traitement Automatique des Langues*, 50-2.

Bases textuelles

ARTFL – *Project for American and French Research on the Treasury of the French Language* (1982-), Chicago. Université de Chicago, *Division of the Humanities and Electronic Text Services*, et ATILF-CNRS, https://artfl-project.uchicago.edu/.

BFM – *Base de Français Médiéval* (2012-), Lyon. UMR ICAR, CNRS, ENS de Lyon, Laboratoire ICAR, http://bfm.ens-lyon.fr.

BVH – *Bibliothèques Virtuelles Humanistes* (2012-), Tours. CESR, Université de Tours, http://www.bvh.univ-tours.fr.

CÉPM – *Corpus Électroniques de la Première Modernité* (2013-), Paris. Université Paris Sorbonne, http://cepm.paris-sorbonne.fr.

Base textuelle Frantext, ATILF - CNRS & Université de Lorraine, http://www.frantext.fr.

MCVF – *Modéliser le changement : les voies du français* (2011-), Ottawa. Université d'Ottawa, http://continent.uottawa.ca/fr/corpus/corpusmcvf/.

NCA – *Nouveau Corpus d'Amsterdam* (2006-), Stuttgart. Université de Stuttgart, institut de linguistique/romanistique, http://www.uni-stuttgart.de/lingrom/stein/corpus.