
Variations langagières et annotation morphosyntaxique du latin classique

Céline Poudat* - Dominique Longrée**

*. « Bases, corpus, langage » (BCL) – UMR6039 – CNRS – Université de Nice
Faculté des lettres
98, bd. Édouard-Herriot,
BP 3209 ; 06204 Nice CEDEX 3
celine.poudat@unice.fr

** Laboratoire d'analyse statistique des langues anciennes (LASLA)
Université de Liège - Bâtiment A4
Quai Roosevelt 1B
B 4000 Liège
dominique.longree@ulg.ac.be

RÉSUMÉ. Le présent article évalue les performances de trois étiqueteurs morphosyntaxiques entraînaibles (MBT, TnT et TreeTagger) pour l'annotation morphosyntaxique de textes latins classiques. Pour ce faire, les textes-tests et les corpus d'entraînement ont été choisis au sein de la banque de données du Laboratoire d'analyse statistique des langues anciennes (LASLA) de manière à mettre les étiqueteurs à l'épreuve de variations stylistiques, diachroniques, génériques ou discursives. Cette recherche a permis non seulement de préciser les divers taux de réussite de chacun des étiqueteurs selon les corpus retenus, mais aussi de montrer que ceux-ci constituent de véritables instruments heuristiques pouvant permettre d'améliorer de manière significative la description des corpus.

ABSTRACT. This paper assesses the performance of three taggers (MBT, TnT and TreeTagger) when used for the morphosyntactic annotation of classical Latin texts. With this aim in view, we selected the training corpora, -as well as the samples used for tests-, from the texts of the LASLA database. The texts were chosen according to their ability to allow testing of the taggers sensitivity to stylistic, diachronic, generic or discursive variations. On the one hand, this research pinpoints the achievements of each tagger according to the various corpora. On the other hand, the paper proves that these taggers can be used as true heuristic instruments and can help to improve significantly the description of the corpus.

MOTS-CLÉS : latin classique, style, genre, discours, morphosyntaxe, étiqueteurs.

KEYWORDS : classical latin, style, genre, discourse, morphosyntax, taggers.

1. Introduction

Au début des années soixante, les philologues classiques se sont intéressés très rapidement aux possibilités nouvelles qu'offraient le développement tout d'abord des techniques mécanographiques, puis des traitements informatiques. Dès 1961, le Laboratoire d'analyse statistique des langues anciennes (LASLA) de l'Université de Liège entreprenait de lemmatiser et d'étiqueter l'ensemble du corpus des textes littéraires classiques grecs et latins¹.

Pour ce faire, le LASLA a mis au point une méthode qui est toujours appliquée aujourd'hui et dont la fiabilité n'est contestée par aucun philologue². Sa particularité majeure est de ne pas se contenter d'une simple lemmatisation des textes, mais d'associer à chaque forme, non seulement son lemme, mais également un code alphanumérique correspondant à une analyse morphosyntaxique, partielle pour les textes grecs, mais complète pour les textes latins. En outre, dans le cas de ces derniers, un code syntaxique permet d'identifier les prédicats des propositions principales et subordonnées, tout en spécifiant la nature des dites subordonnées. Le processus d'encodage repose sur l'utilisation d'un programme informatique : s'appuyant sur un dictionnaire informatisé, ce dernier propose pour chaque forme, tous les lemmes et toutes les analyses possibles. Le choix des analyses pertinentes résulte ensuite d'un travail de sélection manuelle de la part d'un philologue. Cette méthode présente le grand avantage d'offrir des listes d'analyses vérifiées systématiquement, ce qui assure la précision de l'annotation des textes et la validité des études statistiques menées ensuite sur cette annotation³.

Au terme de près de cinquante années d'encodage de textes latins, le LASLA a pu réunir aujourd'hui un corpus lemmatisé et annoté précisément de près de 2 millions de mots. L'essor de l'informatique et des possibilités de numérisation a certes permis la constitution d'autres grands corpus de langues anciennes et en particulier d'œuvres latines, tant sur CD-Rom (PHI5, BTL, etc.) que *via* Internet

1. Cf. (Ét. Évrard 1962).

2. Cette méthode a été décrite à plusieurs reprises dans divers périodiques (Ét. Évrard et A. Bodson, 1966 ; J. Denooz, 1978 ; S. Mellet et G. Purnelle, 2002).

3. à la différence de projets tels que LEMLAT (<http://www.ilc.cnr.it/lemlat/>), Collatinus (<http://collatinus.fltr.ucl.ac.be/>) ou encore Perseus (<http://www.perseus.tufts.edu/hopper/>), qui offrent simplement la liste de toutes les analyses possibles, éventuellement avec une indication préférentielle. Ces outils présentent d'autres inconvénients : par exemple, ils sont généralement incapables de reconnaître un parfait passif dans une forme telle que *amatus est*, mais analysent *amatus* comme participe parfait ; on remarque également que Perseus (<http://www.perseus.tufts.edu/hopper/morph.jsp?l=quarum&la=la&prior=tres>) ne semble pas analyser comme un pronom relatif une forme telle que le *quarum* de la première ligne de la *Guerre des Gaules*...

(*The Latin Library, Perseus Project, Itinera electronica, etc.*)⁴ : tantôt il ne s'agit que de simples numérisations des textes tels qu'ils apparaissent dans les éditions utilisées (dont les provenances sont parfois totalement inconnues), tantôt les textes sont pourvus d'une lemmatisation minimale, parfois aussi sont proposées toutes les analyses possibles d'une partie des formes, avec éventuellement l'indication de la forme la plus probable. Mais la banque de textes latins du LASLA reste la seule, à ce jour, à présenter un volume de textes numérisés à la fois tirés d'éditions clairement identifiées, pourvus d'une annotation morphosyntaxique complète systématiquement contrôlée et constituant un corpus véritablement significatif et représentatif.

La communauté rassemblée autour des corpus latins numérisés se rattachant davantage à une tradition philologique qu'informatique, il s'agissait avant tout pour celle-ci de disposer d'informations précises et sûres sur les textes. Cette exigence a fort probablement ralenti le développement pour le latin d'étiqueteurs morphosyntaxiques ou encore d'assignateurs automatiques de catégories (*taggers*), alors que ceux-ci ont été largement développés ces dernières années pour les langues vivantes et qu'il est communément admis que les résultats qu'ils obtiennent sont tout à fait honorables.

Les choses ont toutefois commencé à évoluer pour le latin, en raison notamment du développement des études médio- et néo-latines, et face à l'importance des corpus à traiter, bien plus grands que ne l'est le corpus latin classique et permettant plus difficilement d'envisager un étiquetage semi-manuel. Dans cette perspective, les possibilités d'entraînement ou d'apprentissage que proposent certains étiqueteurs, sont particulièrement intéressantes : il est ainsi possible de générer un outil d'annotation automatique à partir d'un corpus manuellement étiqueté, le système d'annotation et la langue de départ étant libres. Cette possibilité intéresse particulièrement la communauté linguistique en permettant d'annoter les langues anciennes⁵, ou non prises en charge par les systèmes existants⁶, l'oral, ou encore les genres et les discours⁷

Entraîner un étiqueteur requiert *a minima* un corpus annoté et, dans la mesure du possible, exempt d'erreurs d'annotation. Le coût de constitution d'un tel corpus est donc extrêmement élevé, ce qui explique le fait que la plupart des concepteurs entraînent leurs outils sur les mêmes corpus hétérogènes et déjà annotés (le Penn TreeBank pour l'anglais), afin de générer des outils généralistes, supposés étiqueter « en langue » et s'appliquer à tout texte. Différentes études (D. Biber, 1993 ;

4. (J. Denooz, 2003). Pour plus d'informations, on peut consulter en ligne la *Bibliotheca Classica Selecta* (<http://bcs.fltr.ucl.ac.be/SLInf.html>).

5. Sur ces développements, voir, par exemple, le projet MorphoMultiDim dirigé par S. Heiden pour l'ancien français ou le projet Perseus (D. Bamman *et al.*, 2007).

6. Voir, par exemple, pour le suédois, (J. Sjöbergh, 2003).

7. Voir, par exemple, pour l'annotation du genre scientifique de l'article de linguistique (C. Poudat, 2006).

G. Illouz, 1999) ont pourtant montré que la qualité des sorties variait substantiellement d'un (type de) texte et d'un outil à l'autre : un outil entraîné sur un corpus homogène en genre, par exemple, obtiendrait ainsi de meilleurs résultats sur un texte rattaché au même genre que sur un texte génériquement éloigné.

Partant de ces constats, et fort du corpus de latin classique manuellement annoté du LASLA, nous avons voulu évaluer les étiqueteurs disponibles dans le cadre d'une réflexion plus générale sur les typologies textuelles : est-ce qu'un étiqueteur entraîné sur un auteur ou un discours donné obtient de meilleurs résultats pour annoter un autre texte de ce même auteur ou de ce même discours ? Que peut-on en déduire de l'étiqueteur, ou de l'homogénéité des corpus ?

Cette entreprise est facilitée par le fait que le corpus latin classique est fermé et très bien documenté sur le plan philologique. On évaluera donc les étiqueteurs en les mettant à l'épreuve de variations stylistiques, diachroniques, génériques ou discursives, ce qui nous permettra d'une part de comparer les performances des étiqueteurs et, d'autre part, de participer à la description des corpus.

2. Corpus et méthodes

L'évaluation portera sur les performances de trois étiqueteurs (MBT Tagger, TnT Tagger et le TreeTagger) pour l'annotation de cinq textes-tests : le livre 3 de la *Guerre des Gaules* de César – BGall3 (3 673 tokens), la *Conjuration de Catilina* de Salluste – SalCat. (10 688 tokens), le livre 3 de l'*Histoire d'Alexandre le Grand* de Quinte-Curce, QC3 (7 261 tokens), la *Première Catilinaire* de Cicéron- CicCat1 (3 333 tokens) et le poème 66 de Catulle, Catu66 (586 tokens). La mise en évidence des facteurs de contraste se fera en variant la nature du corpus d'évaluation.

2.1. Trois étiqueteurs sur le banc d'essai

Nous nous sommes concentrés sur trois étiqueteurs : MBT Tagger, TnT Tagger et le TreeTagger, qui illustrent trois techniques d'étiquetage et d'apprentissage différentes : mémoire (MBT), trigrammes (TnT) et n-grammes avec arbres de décision (TreeTagger). On soulignera que nous avons rapidement écarté de notre entreprise l'étiquetage par règles, dont Brill tagger est le représentant le plus illustre : après différents tests, la procédure d'apprentissage s'est révélée beaucoup trop coûteuse. La phase d'entraînement requiert plusieurs étapes, et le programme est écrit en Perl, langage interprété beaucoup plus lent qu'un langage compilé comme C, par exemple. Il faut environ trois jours pour générer les premières règles à partir d'un corpus de 250 000 mots sur une machine de performances honorables. Notre étude étant fondée sur des phrases d'entraînements successifs, nous avons dû écartier le tagger. Notons toutefois que nous envisageons à terme de recourir à des réimplémentations de Brill.

2.1.1. *MBT tagger*

Le tagger MBT (ILK, W. Daelemans) est fondé sur des techniques d'apprentissage et de classification, et concrètement sur le logiciel de classification linguistique TiMBL (également distribué par les mêmes concepteurs).

Les méthodes d'apprentissage fondées sur la mémoire (*memory-based learning*) fonctionnent sur l'idée qu'un raisonnement par analogie est tout aussi efficace, voire plus approprié qu'un raisonnement fondé sur des règles (comme Brill). Dans cette perspective est stocké un ensemble de représentations de situations antérieures (les exemples) destiné au traitement d'éléments nouveaux à partir d'une mesure de similarité.

Générer un tagger s'effectue à l'aide d'un fichier organisé en deux colonnes séparées par une espace : dans la première sont consignés les mots ou les marques de ponctuation, et, dans la seconde, les étiquettes.

La procédure d'entraînement est relativement aisée : MBT génère deux bases à partir du texte d'entrée, une base pour les mots connus et une autre destinée au traitement des mots inconnus. Soulignons que l'utilisateur doit fournir des informations concernant le contexte (contextes droit et gauche, portée) et la forme des mots à étiqueter (position, accents, caractères numériques, etc.) : un certain nombre de tests doit donc être effectué afin de générer le tagger le plus pertinent.

2.1.2. *TnT tagger*

TnT (Trigrams'n'Tags) tagger est un étiqueteur statistique constitué d'un ensemble de méthodes de lissage (*smoothing*)⁸ et de traitement des mots inconnus, implémenté sur un algorithme fondé sur les modèles de Markov (Th. Brants, 2000).

Entraîner TnT nécessite d'enregistrer un corpus étiqueté en format .TT dans le répertoire MODELS et de l'appeler en paramètres de l'étiquetage *via* la commande *tnt-para*.

L'étiqueteur est un modèle génératif qui obtiendrait de meilleurs résultats sur des corpus d'entraînement de taille restreinte que les modèles discriminants (St. Clark *et al.*, 2003) et qui a obtenu les meilleurs taux de précision dans diverses études (J. Zavrel et W. Daelemans, 2000 ; J. Sjobergh, 2003).

2.1.3. *TreeTagger*

TreeTagger (TC Project, H. Schmid) se rapproche des taggers *n-grammes* traditionnels, mais utilise un arbre de décision binaire pour calculer la taille du contexte à utiliser afin d'estimer les probabilités de transition.

8. Ajustement de l'ensemble des données au modèle, ou à la courbe.

Il fonctionne avec deux programmes : *train-tree-tagger*, qui génère un fichier paramètre à partir d'un lexique et d'un corpus manuellement balisé, et *tree-tagger*, qui prend un fichier paramètre et un fichier texte en arguments et qui permet d'étiqueter les textes.

Trois fichiers sont requis pour entraîner l'étiqueteur : un lexique de formes fléchies, un fichier contenant les étiquettes susceptibles d'être affectées aux mots inconnus et le corpus d'entraînement contenant les données étiquetées.

2.2. Corpus de contraste et d'évaluation

Divers corpus d'entraînement seront successivement utilisés de manière à pouvoir mettre en évidence et mesurer l'impact des éventuels facteurs de variation : style de l'œuvre, style de l'auteur, diachronie, genre littéraire ou type de discours.

2.2.1. L'unité stylistique de l'ouvrage à l'épreuve du style de l'auteur

Dans un premier temps, il s'agira d'évaluer si l'on obtient de meilleurs résultats avec un apprentissage limité à l'ouvrage même dont est issu le texte-test, ou si, au contraire, il vaut mieux entraîner l'étiqueteur sur un corpus plus vaste constitué de l'ensemble des œuvres conservées d'un même auteur. Pour ce faire, nous comparerons les résultats obtenus pour le troisième livre de la *Guerre des Gaules* BGall03 selon le corpus d'entraînement choisi : soit le reste de la *Guerre des Gaules*, livres 1-2 et 4-7 (42 055 tokens), soit l'autre œuvre de César conservée, la *Guerre civile* (33 221 tokens), soit le cumul des deux corpus précédents (75 277 tokens). Dans un objectif de comparaison et de contraste, les étiqueteurs entraînés sur ces corpus serviront également à une annotation des autres textes-tests.

2.2.2. Le style personnel de l'auteur à l'épreuve du genre

Ce corpus de César n'est certes pas totalement homogène sur le plan stylistique, mais il correspond toutefois au style personnel d'un seul auteur. Dans un deuxième temps, nous comparerons les résultats obtenus sur ce corpus césarien avec ceux produits par un corpus qui associera des auteurs différents et donc des styles personnels différents. Tous les textes appartiendront toutefois au même genre littéraire, en l'occurrence le genre historique, et à la même synchronie, à savoir le milieu du premier siècle avant J.-C., soit la fin de l'époque républicaine. Ce nouveau corpus d'entraînement comprendra tous les textes des imitateurs de César (livre 8 de la *Guerre des Gaules*, *Guerre d'Afrique*, *Guerre d'Alexandrie*, *Guerre d'Espagne*), ainsi que la *Guerre de Jugurtha* de Salluste (au total 65 532 tokens). On additionnera ensuite ce nouveau corpus au corpus césarien (au total 137 809 tokens). Il s'agira de voir notamment dans quelle mesure ces deux modifications apportées au corpus d'entraînement affectent les résultats pour BGall03 et si elles améliorent bien, comme on pourrait s'y attendre, les

performances obtenues pour l'autre œuvre de Salluste, à savoir la *Conjuration de Catilina* - SalCat.

2.2.3. *Le genre à l'épreuve du temps*

Restant au sein du genre historique, on élargira diachroniquement le corpus d'entraînement : on constituera tout d'abord un corpus d'entraînement composé des livres 4 à 10 de l'*Histoire d'Alexandre le Grand* de Quinte-Curce et de l'ensemble de l'œuvre historique de Tacite, soit la *Vie d'Agricola*, les *Histoires* et les *Annales* (21 5011 tokens) ; on additionnera ensuite ce corpus au corpus historique défini en 2.2.2. Ce nouveau corpus (352 820 tokens) couvrira plus d'un siècle et demi, la *Guerre des Gaules* pouvant être datée de la moitié du premier siècle avant J.-C., alors que l'*Histoire d'Alexandre le Grand* date vraisemblablement du milieu du premier siècle après J.-C. et que la rédaction des œuvres de Tacite se situe entre 96 et sans doute 120 après J.-C. On tentera de préciser si travailler sur l'ensemble du genre apporte ou non de meilleurs résultats pour les textes-tests relevant de ce genre et, si ce n'est pas le cas pour tous, éventuellement pour lesquels d'entre eux (on s'attend bien évidemment à une amélioration des résultats pour le livre 3 de l'*Histoire d'Alexandre le Grand* de Quinte-Curce).

2.2.4. *La prose à l'épreuve des variations génériques : histoire vs. discours vs. dialogue*

On testera ensuite le poids des variations génériques en constituant de nouveaux corpus d'apprentissage à partir de textes datant de la même époque que les textes du corpus historique républicain défini en 2.2.2., mais appartenant à deux genres différents : d'un part l'ensemble des discours de Cicéron – à l'exception bien évidemment du texte-test que constitue la *Première Catilinaire* de Cicéron - CicCat1 (417 245 tokens), d'autre part trois dialogues/traités philosophiques du même auteur, le *De Senectute*, le *De Amicitia* et le *De Officiis* (52 252 tokens). On réunira ces deux corpus de textes dus au même auteur (469 497 tokens), pour les additionner, dans une dernière étape, avec le corpus historique républicain, et constituer ainsi un vaste corpus (607 306 tokens) de textes contemporains en prose. On espère ainsi pouvoir évaluer avec précision le poids du genre dans les performances obtenues par les étiqueteurs.

2.2.5. *Le latin républicain à l'épreuve des variations discursives : prose vs. vers*

Enfin, reprenant ce dernier corpus d'entraînement, on y ajoutera encore l'ensemble des poésies de Catulle (à l'exception du texte-test, le poème 66 – Catu66). Le choix de ce poète (12 402 tokens) s'explique pour plusieurs raisons : d'une part, celui-ci est un contemporain de César et de Cicéron, d'autre part son œuvre a servi de modèle à la poésie élégiaque latine dont on peut penser qu'elle est plus éloignée de la prose que ne l'est, par exemple, la poésie didactique d'un

Lucrèce, autre poète contemporain de César et Cicéron dont nous aurions pu également retenir l'œuvre. En insérant l'œuvre de Catulle dans le corpus d'entraînement (qui atteint ainsi 619 708 tokens), nous espérons pouvoir évaluer le poids de l'opposition discursive entre prose et vers sur les performances des étiqueteurs.

2.3. *Système d'étiquetage*

La structure des fichiers du LASLA a demandé quelques adaptations avant que ceux-ci ne puissent servir à l'entraînement des étiqueteurs. La structure traditionnelle de ces fichiers est la suivante :

1. le lemme, tel qu'il figure dans le dictionnaire choisi comme ouvrage de référence⁹ ;
2. un indice permettant de distinguer différents lemmes homographes ou de marquer les noms propres et les adjectifs qui en dérivent ;
3. la forme telle qu'elle apparaît dans le texte (le mot) ;
4. la référence conforme aux règles de l'ars citandi ;
5. l'analyse morphologique complète sous un format alphanumérique ; c'est-à-dire pour un substantif, la déclinaison, le cas et le nombre, pour un verbe, la conjugaison, la voix, le mode, le temps, la personne et le nombre ;
6. pour les verbes, des indications syntaxiques ; les propositions principales sont distinguées des subordonnées, lesquelles sont classées par type de subordonnants.

9. Le choix s'est porté sur le *Lexicon totius latinitatis* de Forcellini (éd. de Corradini, Padoue, 1864-1887 ; réimpr. 1940, avec additions de G. Perin).

Les fichiers se présentent donc comme suit :

1. Lemme	2.	3. Forme du texte	4. Référence	5. Morpho.	6.
VRBS		URBEM	41 001 0001 001 001	13C00	
ROMA	N	ROMAM	41 001 0001 002 002	11C00	
AB		A	41 001 0001 003 003	70600	
PRINCIPIVM		PRINCIPIO	41 001 0001 004 004	12F00	
REX		REGES	41 001 0001 005 005	13J00	
HABEO		HABUERE	41 001 0001 006 006	52L14	&
LIBERTAS		LIBERTATEM	41 001 0001 007 007	13C00	
ET	2	ET	41 001 0001 008 008	81000	
CONSVLATVS		CONSULATUM	41 001 0001 009 009	14C00	
LVCIVS	N	L.	41 001 0001 010 010	12A00	
BRVTVS	N	BRUTUS	41 001 0001 011 011	12A00	
INSTITVO		INSTITUIT	41 001 0001 012 012	53C14	&

Dans l'extrait de fichier qui précède, pour *urbem*, l'analyse 13C signifie que ce mot est un substantif (1) de la troisième déclinaison (3) à l'accusatif singulier (C). La forme verbale *habuere* est analysée 52L14, ce qui signifie verbe (5) de la deuxième conjugaison à la voix active (2), troisième personne du pluriel (L), indicatif (1) parfait (4). Le signe & signifie que *habuere* est un verbe de proposition principale. Dans certains cas, le nombre d'informations pour une forme atteint la dizaine. Ainsi pour une forme telle que *regnante*, on indique les données de référence, le lemme, la catégorie grammaticale, la conjugaison, la voix, le cas, le nombre, le mode, le temps et le genre. Enfin, on précisera éventuellement qu'il s'agit du verbe d'un ablatif absolu.

Les informations fournies par les fichiers du LASLA présentent donc une granularité très fine dont on pouvait présager qu'elle allait compliquer la tâche des étiqueteurs. Nous avons donc décidé, pour ce premier test, de ne prendre en compte ni les lemmes (qui ne pouvaient pas être traités ni par TnT, ni par MBT, mais seulement par TreeTagger dont nous avons dû nous séparer pour d'autres raisons ; cf. *infra* 2.4.), ni les annotations syntaxiques, et de nous contenter des indices de désambiguïsation (colonne 2) et du code correspondant à l'analyse morphologique (colonne 5). Malgré cette réduction des annotations prises en compte, la granularité du tagset demeure très fine et le nombre de tags rencontrés dans les corpus d'entraînement a été encore particulièrement élevé : 3 732 tags au total.

2.4. Synthèse de l'entraînement, ou comment nous fûmes contraints de nous séparer du TreeTagger

L'entraînement de TnT Tagger et de MBT Tagger s'est révélé plutôt aisé. TnT Tagger ayant été optimisé en temps d'étiquetage et d'entraînement, il s'est avéré remarquablement souple d'utilisation, et très rapide à entraîner (moins d'une minute, voire quelques secondes suffisent pour un corpus constitué de plusieurs centaines de textes).

L'entraînement de MBT Tagger s'est avéré également rapide, comme le montre le tableau qui suit.

	TEMPS sec.	Minutes	Mots/sec.
1a. BGall	26	0,43	3 239
1b. BCiv	22	0,37	3 023
1c. César	45	0,75	3 349
2a. Imitateurs et Salluste	39	0,65	3 212
2b. Histoire 1er s. a.C.n.	81	1,35	3 407
3a. Q.-Curce – Tacite	130	2,17	3 314
3b. Genre_historique	205	3,42	3 448
4a. Discours Cicéron	235	3,92	3 581
4b. Traités Cicéron	40	0,67	2 625
4c. Cicéron	268	4,47	3 532
4d. Prose	341	5,68	3 585
5a. Poésie_Catulle	10	0,17	2 484
5b. Prose + Poésie_Catulle	356	5,93	3 504

Tableau 1. Temps d'entraînement de MBT sur les différents corpus

Si nous fondions bien des espoirs sur le TreeTagger, étant donné qu'il était le seul des trois outils retenus à fournir une lemmatisation, nous avons malheureusement dû renoncer à l'entraîner. En effet, TreeTagger n'est pas conçu pour prendre en charge des systèmes d'étiquetage de granularité élevée. Il est limité à 200 étiquettes¹⁰, ce qui contrarie fortement les systèmes d'annotation des latinistes.

10. Nous remercions Helmut Schmidt de cette information précieuse. Le concepteur de TreeTagger nous a d'ailleurs renvoyés à RFTagger, qui supporte des systèmes d'étiquetage plus complexes, et qui n'est pas limité de la sorte.

2.5. Protocole d'évaluation

Les deux étiqueteurs conservés associent à chaque segment une étiquette morphosyntaxique (correcte ou erronée). L'étiquetage fourni n'étant pas ambigu, nous avons évalué chaque sortie en précision, en la comparant au texte de référence correctement étiqueté qui lui était associé.

La précision de l'étiquetage de chaque texte annoté est ainsi exprimée en pourcentage d'étiquettes assignées correctement par rapport à l'ensemble des étiquettes annotées :

Précision = Nombre d'étiquettes correctement annotées / (Nombre d'étiquettes correctement annotées + Nombre d'étiquettes erronées)

Dans le cadre de la présente étude, nous n'avons pas différencié les étiquettes partiellement erronées des étiquettes intégralement incorrectes, même si la différenciation aurait été pertinente. Nombre des erreurs relevées portent en effet uniquement sur les traits flexionnels, alors que la catégorie grammaticale est correctement assignée : par exemple, pour le poème 66, de Catulle, TnT entraîné sur le corpus BGall + BCiv, analyse erronément *lumina* comme A312, c'est-à-dire comme un nom de la troisième déclinaison au nominatif pluriel et non pas comme A332, c'est-à-dire comme un nom de la troisième déclinaison, mais à l'accusatif pluriel. L'erreur se comprend d'autant mieux que *lumina* peut effectivement être un nominatif pluriel. Dans d'autres cas, l'erreur ne porte que sur la catégorie grammaticale, mais pas sur les traits flexionnels : ainsi pour les mêmes textes, étiqueteur et corpus d'entraînement, on rencontre pour *amantes* l'analyse C5121, soit adjectif de la deuxième classe imparisyllabique au nominatif pluriel positif, alors que l'analyse retenue dans le fichier du LASLA est A312, soit nom de la troisième déclinaison nominatif pluriel. Ici aussi l'analyse proposée par l'étiqueteur de *amantes* comme un adjectif se rencontre au sein des fichiers du LASLA dans d'autres contextes. Les erreurs totales sont parfois explicables : ainsi, toujours avec les mêmes textes, étiqueteur et corpus d'entraînement, *ortus* est analysé B40101433, soit comme une forme verbale au parfait de l'indicatif (= *ortus est*), alors que l'analyse attendue serait A432, soit une forme nominale. L'erreur se comprend ici fort bien dans la mesure où *ortus* substantif n'apparaît qu'une seule fois dans le corpus césarien (sous la forme *ortum*), contre trois occurrences de *ortus* dans des formes verbales de parfait, dont une exactement sous la forme *ortus est*.

Il nous était impossible ici de pondérer les erreurs commises, mais une différenciation entre les types d'erreurs pourrait se révéler fort instructive et fera l'objet d'analyses ultérieures.

3. Contrastes commentés

3.1. L'unité stylistique de l'ouvrage à l'épreuve du style de l'auteur

Étiqueteur	Corpus d'entraînement	BGall3	SalCat	QC3	CicCat1	Catu66
MBT	BGall 1-2 et 4-7	79,64	62,93	60,65	57,19	40,27
	BCivile	75,66	63,53	59,44	56,11	43,86
	BGall + BCiv	82,96	67,96	64,51	61,90	44,88
TnT	BGall 1-2 et 4-7	84,26	67,71	68,09	60,94	44,03
	BCivile	80,15	67,31	66,99	58,15	49,15
	BGall + BCiv	86,33	71,88	71,42	63,73	52,22

Tableau 2. L'unité stylistique de l'ouvrage à l'épreuve du style de l'auteur : pourcentages d'étiquettes correctes

Le premier tableau présente les résultats obtenus en entraînant successivement MBT et TnT sur la *Guerre des Gaules* moins le texte-test (BGall 1-2 et 4-7), sur la *Guerre civile* (BCivile), puis sur l'ensemble des deux textes. Les deux étiqueteurs réagissent globalement de manière similaire : pour BGall3, les résultats sont meilleurs sur le corpus entraînement constitué par la *Guerre des Gaules* (79,64 % ou 84,26 % de réussite) que par la *Guerre civile* (75,66 % ou 80,15 %). On retrouve ici la preuve de la distance stylistique existant entre les deux œuvres, distance par ailleurs bien connue des latinistes¹¹. Dans les deux cas, les meilleurs résultats sont toutefois obtenus avec le corpus d'entraînement cumulant les deux textes et donc le plus étendu (82,96 % et 86,33 %). Il est par ailleurs assez surprenant de voir que de relativement bons résultats sont encore obtenus pour Salluste et Quinte-Curce, principalement d'ailleurs pour TnT (sur l'ensemble de l'œuvre de César, respectivement 71,88 % et 71,42 %). À noter ici que la distance dans le temps séparant le corpus d'apprentissage et l'œuvre de Quinte-Curce ne semble jouer qu'un rôle limité : avec TnT entraîné sur la seule *Guerre des Gaules*, Quinte-Curce obtient même de meilleurs résultats (68,09 %) que le *Catilina* de Salluste (67,71 %), texte de très peu postérieur à la *Guerre des Gaules*. Les moins bons résultats sur la *Première Catilinaire* de Cicéron s'expliquent ici sans doute par l'appartenance de l'œuvre à un genre littéraire différent, le discours, mais, comme on le verra à la lecture du tableau 3, aussi sans doute essentiellement par des différences de style personnel entre César et Cicéron. On relèvera enfin les piètres résultats de l'étiqueteur lorsqu'il s'agit d'un type de discours totalement différent comme la poésie de Catulle. Du point de vue de la comparaison entre étiqueteurs, TnT donne systématiquement des résultats meilleurs que ceux de MBT.

11. Cf. notamment (D. Longrée, 2005).

3.2. Le style personnel de l'auteur à l'épreuve du genre

Étiqueteur	Corpus d'entraînement	BGall3	SalCat	QC3	CicCat1	Catu66
MBT	BGall + BCiv	82,96	67,96	64,51	61,90	44,88
	imitateurs + Salluste	77,48	74,40	67,46	66,85	48,81
	Histoire 1 ^{er} s. a.C.n.	84,45	77,17	70,78	69,64	52,22
TnT	BGall + BCiv	86,33	71,88	71,42	63,73	52,22
	imitateurs + Salluste	82,38	79,04	74,84	72,49	53,58
	Histoire 1 ^{er} s. a.C.n.	87,07	81,28	77,23	70,69	55,80

Tableau 3. *Le style personnel de l'auteur à l'épreuve du genre : pourcentages d'étiquettes correctes*

Dans le tableau 3 on retrouve, aux lignes 1 et 4, les résultats déjà présentés dans le tableau 2 (lignes 3 et 6), obtenus avec l'ensemble de l'œuvre de César. Les lignes 2 et 5 présentent ceux produits par le corpus d'entraînement comprenant les imitateurs de César, ainsi que la *Guerre de Jugurtha* de Salluste. Les lignes 3 et 6 portent sur le corpus d'entraînement regroupant l'ensemble de ces textes : ceux-ci appartiennent tous au même genre littéraire historique, et sont tous presque contemporains, datant du milieu du premier siècle av. J.-C.

On ne s'étonnera pas de voir que les résultats obtenus pour BGall3 sont moins bons sur le corpus d'entraînement se limitant aux ouvrages des imitateurs et de Salluste (MBT : 77,48 % ; TnT : 82,38 %) que sur le corpus d'entraînement formé du reste de l'œuvre de César (MBT : 82,96 % ; TnT : 86,33 %). En revanche, les résultats obtenus pour la *Conjuration de Catilina* de Salluste sont nettement meilleurs avec ce corpus d'entraînement imitateurs et Salluste (MBT : 74,40 % ; TnT : 79,04 %) qu'avec le corpus d'entraînement constitué par l'œuvre de César (MBT : 67,76 % ; TnT : 71,88 %). Cette opposition reflète clairement les différences stylistiques bien connues entre César et Salluste. Par ailleurs, si les scores obtenus pour BGall3 sur le corpus imitateurs et Salluste sont supérieurs à ceux de la *Conjuration de Catilina*, c'est probablement en raison du poids des imitateurs dans ce corpus, la longueur de leurs ouvrages additionnés (44 097 tokens) étant de loin supérieure à celle de la seule *Guerre de Jugurtha* (21 435 tokens). Comme dans le tableau précédent, les meilleurs résultats sont toutefois obtenus avec le corpus d'entraînement le plus grand, et le sont par TnT (87,07 % pour BGall3 et 81,28 % pour Salluste).

Les chiffres obtenus pour Quinte-Curce reflètent bien une proximité plus grande par rapport aux imitateurs et à Salluste que par rapport à César (MBT : 67,46 % contre 64,51 % ; TnT : 74,84 % contre 71,42 %). Les scores sont nettement meilleurs pour Cicéron sur le corpus des imitateurs et de Salluste que sur le corpus

constitué par l'œuvre de César, ce qui tend à confirmer l'opposition stylistique César-Cicéron relevée dans le tableau 2. En outre, on note une première discordance entre les deux étiqueteurs : MBT donne de meilleurs résultats (69,64 %) avec le corpus d'entraînement le plus étendu, alors que TnT remporte les meilleurs scores avec le corpus composé de Salluste et des imitateurs (72,49 %). Le fait peut surprendre quand on sait que le style de Salluste a été souvent perçu comme radicalement opposé à celui de Cicéron. Toutefois on a déjà noté que le poids des imitateurs était fort important dans le corpus d'entraînement : le phénomène relevé ici pourrait essentiellement s'expliquer par une proximité entre les imitateurs et Cicéron (ce qui devrait être vérifié par un entraînement séparé sur les imitateurs et sur Salluste).

3.3. Le genre à l'épreuve du temps

Étiqueteur	Corpus d'entraînement	BGall3	SalCat	QC3	CicCat1	Catu66
MBT	Histoire 1 ^{er} s. a.C.n.	84,45	77,17	70,78	69,64	52,22
	Q.-Curce et Tacite	76,26	76,13	81,31	70,00	58,36
	Genre historique	85,33	81,18	82,41	74,44	60,75
TnT	Histoire 1 ^{er} s. a.C.n.	87,07	81,28	77,23	70,69	55,8
	Q.-Curce et Tacite	82,55	80,62	85,43	74,32	63,31
	Genre historique	87,88	84,66	86,05	77,17	62,80

Tableau 4. Le genre à l'épreuve du temps : pourcentages d'étiquettes correctes

Dans ce tableau, aux lignes 1 et 4, on retrouve à nouveau les résultats déjà présentés dans le tableau précédent (tableau 3, lignes 3 et 6), obtenus avec un corpus d'entraînement reprenant l'ensemble des œuvres historiques écrites vers la moitié du premier siècle av. J.-C. Les lignes 2 et 5 présentent les résultats produits par un corpus d'entraînement comprenant Quinte-Curce et Tacite. Les lignes 3 et 6 portent sur un corpus d'entraînement regroupant l'ensemble de ces textes historiques : ceux-ci appartiennent tous au même genre littéraire, mais couvrent une période de plus d'un siècle et demi.

Les scores obtenus pour César et Salluste sont, comme attendu, moins bons sur le corpus composé de Quinte-Curce et Tacite (lignes 2 et 5) que sur le corpus historique de l'époque républicaine (lignes 1 et 4) : le phénomène est moins net pour Salluste que pour César, ce qui s'explique fort bien quand on sait que les philologues ont régulièrement souligné l'influence de Salluste sur Tacite et quand on se rappelle la proximité de Quinte-Curce avec Salluste déjà mise en évidence dans le tableau 3. À l'inverse, les résultats obtenus pour le livre 3 de l'*Histoire*

d'Alexandre le Grand sont logiquement meilleurs sur le corpus constitué par le reste de l'œuvre de Quinte-Curce et par Tacite. En dehors de l'influence du style de l'œuvre ou de l'auteur, le facteur diachronique semble donc bien jouer ici un rôle, et les résultats reflètent l'évolution du genre historique. Toutefois, les meilleurs scores sont obtenus, à nouveau, sur le corpus le plus étendu reprenant tous les textes historiques (les gains les plus faibles étant réalisés par BGall3) : cette amélioration des scores tend non seulement à confirmer que l'augmentation de la taille du corpus d'apprentissage a généralement un effet bénéfique, mais aussi à démontrer l'existence d'une réelle unité du genre historique.

En ce qui concerne Cicéron, on s'étonne un peu de voir la *Première Catilinaire* obtenir de meilleurs scores avec le corpus Quinte-Curce et Tacite (lignes 2 et 5), qu'avec un corpus de textes contemporains (lignes 1 et 4). Sans doute ce phénomène s'explique-t-il par l'opposition déjà notée précédemment entre César et Cicéron. Les résultats obtenus pour le poème 66 de Catulle sont eux aussi assez surprenants : c'est sur le corpus Quinte-Curce et Tacite, et avec TnT, que ceux-ci sont les meilleurs (63,31 %). Peut-être faut-il y voir une influence de la « couleur poétique » que l'on reconnaît généralement au style de Tacite et que confirme notamment des proximités statistiques de cet auteur avec les poètes (voir notamment la distribution des parties du discours dans la base Latin du CD-Rom Hyperbase-Latin, produit par l'UMR 6039, « Bases, corpus, langage » – CNRS-Université de Nice).

3.4. La prose à l'épreuve des variations génériques : histoire vs. discours vs. dialogue

Étiqueteur	Corpus d'entraînement	BGall3	SalCat	QC3	CicCat1	Catu66
MBT	Histoire 1 ^{er} s. a.C.n	84,45	77,17	70,78	69,64	52,22
	Cicéron Discours	74,98	74,81	69,67	79,51	57,34
	Cicéron Traités	63,79	64,85	58,49	66,16	48,29
	Cicéron	75,61	76,38	70,31	80,80	57,17
	Prose républicaine	84,59	81,31	75,15	81,40	59,04
TnT	Histoire 1 ^{er} s. a.C.n	87,07	81,28	77,23	70,69	55,8
	Cicéron Discours	80,75	80,41	76,85	84,49	63,82
	Cicéron Traités	69,89	70,13	65,98	70,78	50,51
	Cicéron	81,24	81,29	77,52	84,58	63,99
	Prose républicaine	88,08	85,23	81,56	85,18	63,65

Tableau 5. La prose à l'épreuve des variations génériques : pourcentages d'étiquettes correctes

Dans ce quatrième tableau, les résultats obtenus à partir de l'ensemble des textes historiques d'époque républicaine sont comparés successivement à ceux obtenus avec l'ensemble des discours de Cicéron, avec trois de ses traités philosophiques (pour rappel le *De senectute*, le *De amicitia* et le *De officiis*), avec l'addition des ces deux corpus cicéroniens, et finalement avec le cumul de l'ensemble de ces textes, prose cicéronienne et histoire confondue.

Pour BGall3, Catilina et Q-Curce3, les résultats sont en grande partie parallèles, et ce avec les deux étiqueteurs : les scores sont meilleurs avec les discours (lignes 2 et 7) qu'avec les traités (lignes 3 et 8), ce qui pourrait s'expliquer tant par la présence de discours dans les œuvres historiques que de narrations au sein des discours de Cicéron ; les meilleurs scores sont atteints toutefois avec le corpus cicéronien tout entier (lignes 4 et 9). Quand on compare ces derniers scores avec ceux obtenus sur le corpus historique (lignes 1 et 5), on constate toutefois une différence importante entre César d'une part, Salluste et Quinte-Curce d'autre part. Pour César, les scores obtenus avec Cicéron (lignes 4 et 9) sont nettement moins bons que ceux obtenus avec le corpus historique (lignes 1 et 6), alors que, pour Salluste et Quinte-Curce, ceux-ci sont soit à peu près équivalents (avec MBT), soit légèrement meilleurs (avec TnT). Ce résultat peut paraître pour le moins paradoxal quand on sait qu'une œuvre de Salluste fait partie du corpus historique. À l'épreuve des étiqueteurs, les styles de Salluste et de Quinte-Curce paraissent donc aussi proches (ou éloignés, selon le point de vue adopté) de celui de Cicéron qu'ils ne le sont de ceux de César ou de ses imitateurs.

Les scores obtenus pour la *Première Catilinaire* sont bien évidemment meilleurs sur les autres discours de Cicéron (lignes 2 et 7) que sur les traités (lignes 3 et 8) et le gain résultant du cumul des deux corpus cicéroniens n'est que limité (lignes 4 et 9). Pour les quatre textes-tests en prose, les meilleurs scores sont obtenus sur le corpus cumulant l'ensemble des ouvrages en prose.

En ce qui concerne le poème de Catulle, on constate que celui-ci présente des affinités évidentes avec le corpus des discours de Cicéron (lignes 2 et 7), sans doute de par la présence de destinataires explicites tant pour les poèmes de l'un que pour les discours de l'autre. Une différence apparaît ici entre les deux étiqueteurs : MBT obtient le meilleur résultat sur le corpus le plus étendu, alors que TnT les obtient avec le seul corpus cicéronien. Une différence similaire avait déjà été relevée dans le tableau 4, ce qui pourrait donner à penser que TnT est plus sensible que MBT à des facteurs stylistiques ou génériques.

3.5. Le latin républicain à l'épreuve des variations discursives : prose vs. vers

Étiqueteur	Corpus d'entraînement	BGall3	SalCat	QC3	CicCat1	Catu66
MBT	Prose républicaine	84,59	81,31	75,15	81,40	59,04
	Catulle	47,21	49,22	47,58	53,11	54,61
	Prose rép. + Catulle	84,43	81,39	75,68	81,16	62,97
TnT	Prose républicaine	88,08	85,23	81,56	85,18	63,65
	Catulle	52,65	54,78	54,46	57,79	59,56
	Prose rép. + Catulle	88,18	85,25	81,68	85,30	68,09

Tableau 6. *Le latin républicain à l'épreuve des variations discursives : pourcentages d'étiquettes correctes*

Ce sixième et dernier tableau permet d'évaluer l'impact d'une variation discursive majeure, en l'occurrence le passage de la prose à la poésie. On n'est guère surpris de constater les très mauvais résultats obtenus pour tous les textes de prose avec les étiqueteurs entraînés sur les seuls poèmes de Catulle (lignes 2 et 4). Ceux-ci ne tiennent pas seulement à l'opposition discursive entre prose et vers, mais sans doute aussi à la très petite taille du corpus d'entraînement (pour rappel, seulement 12 402 tokens), comme le démontrent les scores obtenus pour le poème 66, moins bons qu'avec le corpus de prose (lignes 1 et 3). Pour ce qui est des textes-tests en prose, l'impact de l'ajout des poèmes de Catulle au corpus de prose républicaine diffère à peine selon l'étiqueteur : avec TnT, tous les scores s'améliorent très légèrement ; avec MBT, ceux-ci se dégradent un peu dans les cas de BGall3 et de CicCat1. En ce qui concerne le poème 66 de Catulle, c'est avec TnT et avec l'ensemble du corpus que l'on obtient les meilleurs scores, ceux-ci restant toutefois de loin inférieurs à ceux de la prose, ce qui s'explique bien par la part limitée des textes en vers dans ce corpus global. D'autres investigations seraient ici nécessaires pour savoir si les meilleurs résultats pour la poésie pourraient être obtenus en grossissant le corpus d'entraînement avec d'autres textes versifiés ou s'il vaudrait mieux constituer un corpus homogène ne comprenant que de la poésie.

4. Conclusions et perspectives

Un premier acquis de cette étude est de permettre d'évaluer avec précision les performances des étiqueteurs testés. Certes, dans un certain nombre de cas, ceux-ci arrivent à obtenir d'assez bons résultats, mais, contrairement à une idée assez répandue parmi ceux qui travaillent sur les textes latins numérisés, il ne paraît guère possible d'espérer obtenir pour l'instant des résultats atteignant 96 ou 97 % de réussite si l'on opte pour une granularité assez fine de l'annotation (près de 4 000 étiquettes différentes pour les analyses du LASLA) et si l'on souhaite

disposer d'une analyse morphosyntaxique complète de chaque forme du texte (à cet égard, au contraire de ce que d'aucuns pensent, Tretagger ne semble pas être la panacée en la matière, puisqu'il s'est révélé incapable de gérer les multiples annotations de la banque de données du LASLA).

Par ailleurs, même si les résultats de MBT sont à l'heure actuelle systématiquement un peu moins bons que ceux de TnT, les deux étiqueteurs réagissent globalement de la même manière et l'augmentation de la taille du corpus d'entraînement va généralement de pair avec une amélioration des résultats obtenus. MBT et TnT réagissent toutefois différemment en ce qui concerne le poème 66 de Catulle : dans ce cas, TnT n'obtient pas nécessairement les meilleurs résultats avec le corpus le plus grand, ce qui pourrait s'expliquer par une plus grande sensibilité de l'étiqueteur aux variations stylistiques, génériques ou diachroniques. Il s'agirait donc de prolonger l'étude et d'examiner de plus près les réactions des étiqueteurs à un corpus d'entraînement incluant un volume important de poésie. On a noté que l'intégration au corpus de prose du petit corpus poétique constitué par l'œuvre de Catulle amenait à des baisses de performance de MBT sur certains textes-tests (BGall3 et CicCat1). On peut se demander si l'augmentation de la part des textes versifiés ne renforcerait pas cette tendance. D'une manière plus générale, pour préciser encore les choses, on cherchera à faire varier plus largement les corpus d'entraînement et à les enrichir à nouveau, ce que la banque de données du LASLA permettra aisément (pour l'instant, seul un tiers environ de la base a été utilisé, sans compter les possibilités d'application à deux autres corpus de textes dont dispose le LASLA, pour le latin médiéval et pour le néolatin¹²). On testera aussi d'autres étiqueteurs, notamment BrillTagger et RFTagger, que nous avons du malheureusement écarter momentanément jusqu'ici pour des raisons pratiques.

La recherche a également souligné l'intérêt des étiqueteurs comme instruments heuristiques. Certes, la comparaison des résultats que ceux-ci obtiennent sur divers textes-tests en fonction des variations des corpus d'entraînement ne remplace pas le recours aux méthodes éprouvées des études philologiques ou aux analyses statistiques de la textométrie, mais celle-ci les complète utilement. Ainsi, malgré leurs différences de performance, MBT et TnT permettent globalement de relever les mêmes tendances : homogénéité du style historique au-delà de l'évolution diachronique, écart entre narration et discours, mais aussi plus spécifiquement entre les styles de César et de Cicéron, écart moins grand entre Catulle et Cicéron ou entre Catulle et Quite-Curce ou Tacite qu'entre Catulle et César, etc. Certaines des informations ainsi recueillies ne sont pas nouvelles, mais le philologue s'étonnera sans doute de constater que la langue de Cicéron semble plus éloignée de celle de César qu'elle ne l'est de celle des imitateurs de ce dernier et de Salluste. L'utilisation des étiqueteurs à des fins heuristiques mériterait donc sans doute d'être approfondie et pourrait non seulement venir conforter d'autres méthodes d'analyses, mais aussi révéler des phénomènes qui ont jusqu'ici échappé aux investigations des latinistes.

12. (D. Longrée et C. Poudat, 2009a et 2009b).

5. Bibliographie

- Bamman D., Passarotti M., Crane Gr. & Raynaud S., « A collaborative model of treebank development », Bergen, 2007, p. 1-6.
- Biber D., « Using register-diversified corpora for general language studies », *Computational Linguistics*, 19(2), 1993, p. 243-258.
- Brants Th., « TnT – A Statistical Part-of-Speech Tagger », in *Proceedings of the 6th Applied NLP Conference*, ANLP-2000, April 29-May 3, 2000, Seattle, WA.
- Clark St., Curran J. et Osborn M., « Bootstrapping POS Taggers using Unlabelled Data », in Daelemans W. et Osborne M. (eds.), *Proceedings of CoNLL-2003*, Edmonton, 2003, p. 49-55.
- Denooz J., « L'ordinateur et le latin, Techniques et méthodes », *Revue de l'organisation internationale pour l'étude des langues anciennes par ordinateur*, 1978, 4, p. 1-36.
- Denooz J., « Littératures classiques et banques de données », *Informatica e scienze umane, Mezzo secolo di studi e ricerche. Lessico Intellettuale europeo Leo S, Olschki*, 2003, p. 107-128.
- Dzeroski S. et al., « Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets », in *Proceedings of Second International Conference on Language Resources and Evaluation, LREC 2000*, p. 1099-1104.
- Évrard Ét., « Le laboratoire d'analyse statistique des langues anciennes de l'Université de Liège », *Mouvement scientifique en Belgique*, 9, 1962, p. 163-169.
- Évrard Ét., Bodson A., « Le programme d'analyse automatique du latin », *Revue de l'organisation internationale pour l'étude des langues anciennes par ordinateur*, 1966, 2, p. 17-46.
- Illouz G., « Méta-Étiqueteur Adaptatif : Vers une utilisation pragmatique des ressources linguistiques », in *Actes TALN'99*, 1999.
- Longrée D et Poudat C., « Linguistic taggers and syntactic parsers for Classical and post-Classical Latin : the new projects of the LASLA », in P. Anreiter et M. Kienpointner (eds.), *Proceedings of the 15th International Colloquium on Latin Linguistics* (Innsbrucker Beiträge zur Sprachwissenschaft), à paraître.
- Longrée D., « Temps verbaux et spécificités stylistiques chez les historiens latins : sur les méthodes d'analyse statistique d'un corpus lemmatisé », in G. Calboli (éd.), *Papers on Grammar, IX, 2, Latina Lingua !, Proceedings of the Twelfth International Colloquium on Latin Linguistics (Bologna, 9-14 June 2003)*, Rome, 2005, p. 863-875.
- Longrée D. et Poudat C., « Latin du Haut Moyen Âge et annotation morphosyntaxique automatique : quelles perspectives ? », in F. Biville, *Actes du IX^e Colloque international « Latin vulgaire – Latin tardif » (LVLT9)*, Lyon, 2-6 septembre 2009, à paraître.
- Mellet S. et Purnelle G., « Les atouts multiples de la lemmatisation : l'exemple du latin », in A. Morin A. & P. Sébillot (eds.), *JADT 2002, 6^{es} Journées internationales d'Analyse statistique des Données Textuelles*, Saint-Malo, INRIA-IRISA, 2002, p. 529-538.

- Poudat C., *Étude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*, Thèse, Orléans, 2006. Disponible sur <<http://www.revue-texto.net/Corpus/Publications/Poudat/Etude.html>>.
- Sjobergh J., « Combining POS-taggers for improved accuracy on Swedish text », in *Proceeding of NoDaLiDa 2003*, Reykjavik, 2003.
- Zavrel J. et Daelemans W., « Bootstrapping a tagged corpus through combination of existing heterogeneous taggers », in *Conference on Language Resources and Evaluation (LREC 2000)*, 2000, p. 7-20.