Structures phrastiques et analyse automatique des données morphosyntaxiques : le projet LatSynt

Dominique Longrée [1], Caroline Philippart de Foy [2], Gérald Purnelle [2]

[1] LASLA – Université de Liège et FUSL (Bruxelles)

[2] LASLA – Université de Liège

Résumé

Le projet LatSynt constitue une recherche originale et novatrice sur l'ordre des mots et sur les structures de l'énoncé latin. Elle s'appuie sur les données morphosyntaxiques de la banque textuelle latine du Lasla et sur la linéarité du texte. Elle vise à développer des procédures d'analyse syntaxique automatisées reposant sur l'implémentation de diverses règles d'ordre de mots, à évaluer la pertinence des descriptions linguistiques récentes, et à fournir de nouveaux outils pour la modélisation des structures énonciatives et pour la classification des textes latins. La première phase du projet a pour objet de délimiter les propositions subordonnées et de préciser leur niveau de subordination.

Abstract

The project LatSynt is an original and innovative research on word order and Latin sentence structures. Based on the morphosyntactic data of the Latin textual bank of the Lasla and on the text linearity, this research aims to develop automatic procedures for parsing based on word order rules, to evaluate the relevance of recent linguistic descriptions, and to provide new tools for enuntiative structure modeling and Latin texts classification. The first stage of the project consists in bounding subordinate clauses and in specifying their subordination level.

Key-words: Latin, syntax, topology, labelling, automatic analysis

0. Introduction

Parallèlement aux recherches menées sur les formes, lexèmes et catégories morphosyntaxiques, les travaux dans le domaine de l'ADT tendent de plus en plus régulièrement à prendre en compte les structures syntaxiques, ce qui suppose un codage spécifique des textes. Pour répondre à ce besoin, divers projets de type Treebank ont opté pour un étiquetage syntaxique manuel, complété éventuellement par un encodage au moyen d'un étiqueteur automatique entraînable. On citera ici, entre autres, les projets Nota Bene (Mazziotta, in press) ou Perseus (Bamman and Crane, 2007 ; Bamman et al., 2007). Cette méthode présente plusieurs inconvénients : coût de l'encodage manuel, fiabilité partielle de l'encodage automatique, choix *a priori* d'un cadre linguistique théorique particulier. En vue d'enrichir les informations contenues dans sa banque de textes latins, le Laboratoire d'Analyse Statistique des Langues Anciennes de l'Université de Liège - LASLA (http://www.ulg.ac.be/cipl/bdlasla/) a donc opté pour une démarche différente, qui cherche à allier étroitement analyse des données et étiquetage du corpus : en nous appuyant tant sur les données morphosyntaxiques déjà encodées dans les fichiers du LASLA que sur la linéarité du texte, nous cherchons à développer des procédures d'analyse syntaxique automatisées

reposant notamment sur l'implémentation de diverses règles d'ordre de mots. Nous espérons ainsi pouvoir d'une part évaluer la pertinence des descriptions linguistiques récentes, d'autre part fournir de nouveaux outils pour la modélisation des structures énonciatives et pour la classification des textes latins.

Dans le cadre de la présente contribution, nous rappellerons tout d'abord le contenu et la structure de la base de données du LASLA. Nous indiquerons ensuite les lignes directrices et les objectifs poursuivis par le projet LatSynt. Nous présenterons les procédures mises en œuvre et les résultats déjà obtenus dans le cadre de sa première phase, qui vise à délimiter les propositions subordonnées et à établir leur niveau d'enchâssement syntaxique. Nous préciserons enfin les complémentarités pouvant exister entre notre démarche et les recherches menées par ailleurs dans les domaines de la topologie textuelle et de l'analyse des motifs syntaxiques ou pluridimensionnels.

1. La banque de données et le système d'annotation du LASLA

Depuis sa création en 1961, le LASLA a constitué une vaste banque de données de textes latins dont la qualité est internationalement reconnue : celle-ci est la seule à associer systématiquement chacune des formes du texte à son lemme et à son analyse morphologique complète. La structure traditionnelle des fichiers du LASLA est la suivante :

1. Le lemme, tel qu'il figure dans le dictionnaire choisi comme ouvrage de référence, à savoir le *Lexicon totius latinitatis* de Forcellini ;
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 ;
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, etc. ;
6. Pour les verbes, des indications syntaxiques ; les propositions principales sont distinguées des subordonnées, lesquelles sont codées par type de subordonnants.

1. Lemme	*2.*	*3.Forme*	*4.Référence*	*5. Morpho.*	*6. Synt.*
SVM	1	erant	CE0060001001001	56L12	&
OMNINO		omnino	CE0060001002002	60000	
ITER		itinera	CE0060001003003	13J000	
DVO		duo	CE0060001004004	31J00 5	
QVI	1	quibus	CE0060001005005	46O32 1	
ITER		itineribus	CE0060001006006	13O00	
DOMVS		domo	CE0060001007007	12F00	
EXEO	1	exire	CE0060001008008	56071	
POSSVM	1	possent	CE0060001009009	56L32	– LN

Dans cet extrait de fichier, pour *itinera*, l'analyse 13J signifie que ce mot est un substantif (1) de la 3e déclinaison (3) au nominatif pluriel (J). La forme verbale *erant* est analysée 56L12, ce qui signifie verbe (5) de la conjugaison anomale à la voix active (6), 3e personne du pluriel (L), indicatif (1) imparfait (2). Le signe & signifie que *erant* est un verbe de proposition principale ; le signe – signifie que la forme *possent* (verbe, conjugaison anomale, 3e personnel du pluriel, subjonctif imparfait) est un verbe subordonné ; le code LN indique qu'il est subordonné par un pronom relatif *qui*.

La base du LASLA ne comporte donc qu'une seule donnée à proprement parler syntaxique, à savoir un code de subordination permettant de repérer les prédicats des propositions principales et des propositions subordonnées. Cependant, en s'appuyant sur cette donnée syntaxique et sur les informations morphologiques très complètes qui ont déjà été encodées, on peut mettre au point des procédures d'analyse syntaxique automatisées permettant de surmonter cette difficulté.

2. Le projet *LatSynt*

2.1. Sur le plan méthodologique et linguistique

Le projet *LatSynt* vise à mener une recherche originale et novatrice sur l'ordre des mots et sur les structures de l'énoncé latin : en faisant appel aux données morphosyntaxiques de la base et à la linéarité du texte, nous tentons d'implémenter un certain nombre de règles d'ordre de mots proposées par des descriptions linguistiques récentes (Charpin, 1977 ; Panhuis, 1982 ; Devine and Stephens, 2006 ; Spevak, 2006 ; 2010). Les résultats obtenus nous permettent non seulement de tester la pertinence des procédures d'analyse et le bien-fondé des descriptions linguistiques sur lesquelles elles s'appuient, mais aussi d'améliorer par étapes successives les programmations correspondantes, en vue de réduire le « bruit » généré. Cette démarche fait ainsi progresser sensiblement notre connaissance des interactions existant, en latin classique, entre ordre syntagmatique et structures syntaxiques au sein des énoncés.

2.2. Sur le plan des applications concrètes

La mise au point d'outils informatiques d'analyse syntaxique a des retombées immédiates pour divers projets de recherche visant à la classification et à la segmentation de textes latins classiques et de textes néo-latins : l'augmentation des informations syntaxiques contenues dans la base permet tant de procéder à des calculs de distance traditionnels basés sur la fréquence qu'à des recherches de type topologique visant à une meilleure prise en compte de la linéarité du texte, tant sur le plan microstructurel qu'au niveau macrostructure l.

3. La première phase du projet : un bornage des propositions

La première phase du projet *LatSynt* [1] vise à développer, parmi les procédures d'analyse syntaxique automatisées, celles qui ont pour objet de *délimiter les propositions subordonnées et de préciser leur niveau de subordination (c'est-à-dire d'enchâssement syntaxique)*.

Les informations syntaxiques encodées initialement dans la base du LASLA ne sont pas suffisantes pour discerner immédiatement les contours des propositions et pour pouvoir connaître leur niveau de subordination syntaxique. La chose est particulièrement compliquée dans le cas des propositions subordonnées sans terme introducteur, comme les propositions infinitives ou les propositions participiales (Ablatifs absolus). Ainsi, par exemple, la succession des codes & AD AG, codes spéciaux marquant les prédicats des propositions, correspondra, dans le texte, à une suite formée par un prédicat de proposition principale (&), un prédicat d'Ablatif absolu (AD) et un prédicat de proposition infinitive (AG), mais, sans retour au texte, on ne pourra en aucune façon savoir si l'Ablatif absolu (AD) est enchâssé dans une proposition infinitive (AG) dont il dépend ou si c'est la proposition infinitive qui dépend de l'Ablatif absolu. Dans

[1] Cette première phase est financée par un Crédit de démarrage de l'Université de Liège et par un Crédit du Fonds de la Recherche Fondamentale Collective du FRS-FNRS de la Communauté française de Belgique.

le cas d'une chaîne & AD AD, la question se pose de manière sensiblement différente : on ne peut déterminer, dans ce cas, s'il s'agit de deux Ablatifs absolus coordonnés ou juxtaposés dépendant tous deux de la proposition principale ou si l'on a affaire à un premier Ablatif absolu enchâssé dans l'autre, ou encore à un second Ablatif absolu dépendant du premier.

Dans le cas des propositions subordonnées introduites par un terme introducteur, les choses sont plus aisées : les créateurs de la base de données du LASLA ont en effet pris soin de prévoir un code de rappel indiquant le mode et le temps du verbe introduit. Dès lors, dans une chaîne & AD LN où LN représente une proposition relative, la prise en compte de la position du relatif permet de savoir si la relative (LN) se rapporte à un élément de l'Ablatif absolu (AD) ou si l'Ablatif absolu est intégré au sein de la relative dont il dépend. De même, dans une suite du type BN BN BN, on pourra déterminer si cette séquence correspond à une seule proposition introduite par la conjonction *cum* incluant trois prédicats coordonnés en une succession rapide ou trois propositions introduites par *cum*, passablement développées et présentant chacune la répétition du terme introducteur : même si la différence est peu importante sur le plan du signifié, celle-ci peut avoir des répercussions non négligeables pour la caractérisation du style ou du genre du texte considéré.

En tenant compte de ces données, nous avons choisi de commencer par la délimitation des propositions à verbes personnels et pourvues d'un terme introducteur. Nous avons donc laissé de côté pour l'instant le bornage des propositions infinitives, des subjonctifs paratactiques, des Ablatifs absolus et des Praedicativa. Voici un aperçu de la procédure de bornage automatique mise au point.

Celle-ci s'appuie sur les informations contenues dans les fichiers de type LASLA d'origine, qui comprennent la notation, en regard de chaque verbe subordonné, d'un code indiquant son type de subordination ou le type de subordonnant qui l'introduit (par exemple AG pour la proposition infinitive, BN pour la proposition introduite par la conjonction *cum* ou LN pour la relative introduite par *qui*) ; en outre, en regard de chaque subordonnant figure la mention du mode et du temps du verbe que celui-ci introduit ou du premier verbe que celui-ci introduit, s'il en introduit une série.

La première étape, préparatoire au bornage automatique, consiste à associer ces deux informations, c'est-à-dire à les faire figurer dans l'enregistrement de tout subordonnant et de tout subordonné ; à l'issue de cette opération, dans une structure telle que, par exemple, *quem […] uidi*, les codes LN14 (signifiant « subordination en QVI » et « indicatif parfait ») sont reportés à la fois dans l'enregistrement de la forme *quem* (lemme QVI) et de la forme *uidi* (lemme VIDEO). Ceci permet à l'étape suivante le repérage des subordonnants et subordonnés potentiellement liés.

La deuxième étape consiste en un programme qui, pour chaque phrase, repère les formes attestées qui présentent de tels codages, c'est-à-dire extrait précisément ces codages et produit un schéma linéaire (de surface) de la structure syntaxique propositionnelle de la phrase. Exemple : &0014 +LN14 -LN14 +LN12 +GK32 -GK32 -LN12, où « & » indique le verbe principal, « + » un subordonnant et « - » un verbe subordonné. À ce stade, on le rappelle, les subordinations sans subordonnants (subjonctifs paratactiques, Ablatifs absolus et propositions infinitives) ne sont pas prises en compte, leur traitement (bornage automatique) étant reporté à une phase ultérieure, plus fine et complexe, de l'exploration des phrases.

Le programme de la troisième étape analyse chaque schéma ainsi produit (pour chaque phrase), déduit les liens potentiels entre tout subordonnant et tout subordonné de même codage, et

ajoute au schéma des signes (crochets et accolades) qui indiquent à la fois ces liens, les bornes théoriques des propositions et, le cas échéant, leur subordination relative (inclusion) :

<&0014>[+LN14 -LN14]{+LN12 [+GK32 -GK32] -LN12}.

Enfin un dernier programme reporte ce codage-bornage dans la phrase proprement dite et la soumet (soumet toutes les phrases) au linguiste :

Tacite, *Annales*, 13,11,2 / <&0014>[+LN14-LN14]{+LN12[+GK32-GK32]-LN12}
<&secuta (est)> que lenitas in Plautium Lateranum [+quem ob adulterium Messalinae ordine demotum -reddidit] senatui clementiam suam obstringens crebris orationibus {+quas Seneca testificando [+quam honesta -praeciperet] uel iactandi ingenii uoce principis -uulgabat}

Il reste alors au philologue à vérifier tout d'abord la pertinence des liens établis entre subordonnant et prédicat, ensuite à vérifier si des éléments appartenant à la subordonnée ne se retrouvent pas soit à gauche du subordonnant, c'est-à-dire en prolepse, soit à droite du prédicat, c'est-à-dire en postposition. C'est ce que nous avons fait sur un corpus comprenant d'une part la *Guerre des Gaules* de César, d'autre part les livres 11 à 16 des *Annales* de Tacite, en nous attachant pour commencer spécifiquement au cas de la prolepse.

4. Les résultats obtenus et les perspectives de recherche

L'analyse automatisée nous a permis de mettre en évidence un certain nombre de régularités portant tout d'abord sur les types de propositions autorisant la prolepse, ensuite sur la nature des syntagmes pouvant apparaître en prolepse et sur leurs fonctions. Nous avons également pu souligner l'obligation pour certains termes, comme le relatif de liaison, d'apparaître en prolepse. Les données recueillies nous ont amenés à étudier la place des cas de prolepses dans la phrase ou dans le texte. Nous avons enfin mis en évidence les écarts qui existent en la matière entre les œuvres des deux historiens.

Nous avons ainsi pu constater que les cas de prolepses étaient assez nombreux devant la conjonction *cum* et, dans une moindre mesure, devant les conjonctions *ut*, *ubi* et *si*. Il s'agit pour une large part de relatifs de liaison, en raison de leur fonction anaphorique, de démonstratifs ayant la même valeur de rappel, de substantifs (accompagnés ou non de relatifs de liaison ou de démonstratifs) et d'adverbes (parmi lesquels principalement *eo*), que l'on rencontre généralement en début de phrase. Ces mots antéposés sont le plus souvent compléments directs ou sujets du verbe de la subordonnée, mais il y a aussi des compléments prépositionnels, des compléments d'un nom de la subordonnée, etc.

L'étude de la répartition des prolepses par rapport aux subordonnants nous a permis de souligner le rapport entre certains termes et certains subordonnants. Les relatifs de liaison, par exemple, se rencontrent principalement devant la conjonction *cum*, puis devant d'autres subordonnées temporelles introduites par *ubi* ou *postquam* ; l'antéposition de l'adverbe *eo* se trouve toujours, dans notre corpus, devant *cum*, à une exception près ; etc. Et, dans l'autre sens, la conjonction *ubi*, par exemple, ne semble admettre comme prolepse, du moins d'après nos relevés, que des relatifs de liaison.

La comparaison des relevés des prolepses dans les deux œuvres a montré un emploi beaucoup plus important chez César que chez Tacite, mais elle nous a également permis de mettre en évidence des différences significatives dans la manière d'écrire des deux auteurs. Par exemple, et pour reprendre des éléments déjà mentionnés, tous les cas de prolepses devant la conjonction *cum* relèvent de César, nous n'en avons rencontré aucun chez Tacite. De même, nous avons pu

noter que, parmi les substantifs antéposés, aucun chez Tacite n'était accompagné d'un adjectif relatif ou démonstratif, contrairement à ce qui se passe chez César. Un examen plus approfondi, tenant compte à la fois de la nature des prolepses et de leur fonction, accompagné d'un retour au texte, nous a également permis de constater une différence dans la manière d'exprimer des choses comparables : la plus grande partie des relatifs antéposés chez Tacite sont des sujets neutres dans des expressions du type *quod ubi auditum est* ou *quae ubi cognita sunt*, tandis que César recourt davantage à l'accusatif pour les relatifs et que l'expression prend chez lui un tour actif (*quod ubi ... animaduertit* ou au masculin *quos cum ... conspexisset*). Enfin, et pour ajouter un nouvel exemple, nous avons pu relever chez César plusieurs conditionnelles placées en prolepse, surtout devant des subordonnées introduites par les conjonctions *ut* et *ne*, mais nous n'en avons trouvé aucune chez Tacite.

Ces premiers résultats, déjà fort significatifs, nous encouragent à poursuivre le processus de bornage des propositions. Mais la recherche va rapidement se compliquer quand il s'agira de déterminer les limites des propositions sans terme introducteur évoquées précédemment. On pourra partir du principe qu'en latin les bornes de celles-ci sont fréquemment marquées par leur sujet et leur prédicat. Dans le cas de la proposition Ablatif absolu, nous rechercherons un nom ou un pronom à l'ablatif se trouvant dans l'environnement du participe et présentant les mêmes genre et nombre que celui-ci. On peut espérer que, dans la plupart des cas, ce nom ou pronom fonctionnera comme sujet du participe et que les deux termes marqueront les bornes extérieures de la proposition : il s'agira de préciser quels autres paramètres devront être employés pour affiner ces premiers critères d'analyse. Dans le cas des propositions infinitives, les choses seront beaucoup plus compliquées : on ne pourra pas en effet s'appuyer systématiquement sur un accord entre le sujet et l'infinitif, puisque seuls les infinitifs futur, potentiel et irréel actifs et les infinitifs parfaits passifs et déponents impliquent un accord. Il s'agira dès lors de repérer un accusatif qui, dans l'environnement de l'infinitif, sera susceptible de fonctionner comme son sujet. Cette démarche impliquera de distinguer l'accusatif sujet d'un éventuel accusatif complément direct de verbe. L'ordre dans la chaîne linéaire devrait ici se révéler déterminant : d'après notamment les travaux de F. Charpin [2], le complément direct ne devrait pas être séparé du verbe dont il dépend par plus d'un syntagme fléchi, alors que l'accusatif sujet peut, lui, en être largement éloigné ; le développement de procédures de repérage automatique offrira une possibilité de vérifier la pertinence de cette description. Par ailleurs, le syntagme sujet pourra se révéler complexe : ainsi, un nom sujet pourra fort bien être précédé par un complément du nom au génitif. Avant d'arriver à borner parfaitement de telles propositions, il faudra donc probablement tenter de procéder au bornage des syntagmes complexes, ce qui constituera la deuxième phase du projet.

Ces diverses recherches permettront de déterminer dans quelle mesure les critères reposant sur la morphosyntaxe et sur l'ordre des mots, même s'ils sont indispensables pour identifier la fonction syntaxique exercée par les différents syntagmes, ne sont pas suffisants à eux seuls : d'autres paramètres, principalement sémantiques, doivent à coup sûr être envisagés ; on peut ainsi penser que le trait animé/inanimé ou la valence verbale jouent un rôle non négligeable dans le repérage du sujet de la proposition infinitive ; la recherche que l'on mènera devrait permettre de le vérifier. On déterminera également ainsi quelles informations sémantiques, voire également phonologiques (accentuelles), seraient nécessaires pour aboutir à une meilleure analyse et dans quelle mesure celles-ci pourraient être ajoutées aux informations contenues

[2] Charpin, 1989 : 503-520.

dans le dictionnaire informatisé du LASLA en vue d'une récupération ultérieure dans le cadre d'analyses automatisées.

5. Les complémentarités avec les recherches en matière de topologie textuelle

Le projet *LatSynt* s'intègre dans le cadre d'un ensemble plus vaste de recherches menées par le LASLA autour de la topologie textuelle dans ses aspects micro- et macro- structurels. Tant en aval qu'en amont de l'analyseur, les interactions sont donc nombreuses entre *LatSynt* et ces autres projets, fruits d'une collaboration régulière avec le laboratoire BCL – « Bases, corpus, langage », Université de Nice – CNRS [3].

Avant même d'implémenter des règles d'ordre des syntagmes et dans un but d'efficacité pratique, il s'agit de vérifier si celles-ci s'appliquent de manière suffisamment récurrente pour autoriser leur transcription dans les programmes de l'analyseur et leur vérification sur nos corpus-tests. À cet égard, les ressources du CD-Rom *Hyperbase Textes Latins* (voir http://www.cipl.ulg.ac.be/Lasla/hyperbasel.html) nous sont particulièrement précieuses. Elles nous permettent non seulement de vérifier la probabilité de séquences syntaxiques par une simple recherche sur les codes du LASLA dans les fonctions « concordance » ou « contexte », mais elles nous offrent aussi la possibilité de rechercher les segments spécifiques de chaque auteur ou de chaque texte. On peut ainsi se rendre compte que, dans les Ablatifs absolus, la séquence « substantif à l'ablatif – prédicat à l'ablatif » domine largement chez César, alors que c'est la séquence inverse qui est la plus caractéristique chez Tacite. Par ailleurs, la recherche menée sur les motifs syntaxiques, également en collaboration avec le laboratoire BCL, fournit des indications précieuses sur l'existence de séquences récurrentes dans les textes servant à tester l'analyseur. La notion de « motif », présentée aux JADT 2008 (Longrée et al., 2008 ; Mellet and Longrée, 2009), présente en outre un cadre théorique permettant de décrire les combinaisons d'éléments lexicaux, morphologiques, syntaxiques, prosodiques, etc. que nos tests mettent en évidence.

En amont de l'analyseur, les résultats obtenus sont directement utilisables par ce même programme de recherche sur les motifs syntaxiques. L'analyseur permet en effet de « désambiguïser » des séquences syntaxiques. Ainsi, comme on l'a dit plus haut, dans une chaîne & AD LN, on ignore si la relative (LN) se rapporte à un élément de l'Ablatif absolu (AD) ou si l'Ablatif absolu est intégré au sein de la relative. Les chaînes & AD +LN -LN (soit prédicat de la principale, prédicat de l'Ablatif absolu, relatif, prédicat de la relative) et & +LN AD -LN (soit prédicat de la principale, relatif, prédicat de l'Ablatif absolu, prédicat de la relative) obtenues grâce à l'analyseur permettent de lever cette ambiguïté : dans le premier cas, il y a une forte probabilité pour que la proposition relative dépende de l'Ablatif absolu et fonctionne donc à un deuxième niveau de subordination, alors que, dans le deuxième cas, ce sera l'Ablatif absolu qui, étant enchâssé dans la relative, en dépendra. Dans une perspective de classification ou de segmentation des textes, cette distinction est importante : la séquence & AD +LN -LN correspond à un « motif de rallonge » particulièrement caractéristique du style de Tacite et dont d'autres recherches (Longrée and Mellet, 2009 ; 2010a ; 2010b ; Longrée et al., 2010) ont montré qu'il pouvait non seulement servir de critère pour distinguer les textes de Tacite

[3] Avec le soutien de Wallonie-Bruxelles International et du Fonds de la Recherche Scientifique, du Ministère Français des Affaires étrangères et européennes, du Ministère de l'Enseignement supérieur et de la Recherche dans le cadre des Partenariats Hubert Curien.

de ceux des autres historiens, mais aussi constituer un marqueur de la structuration du récit, en caractérisant clairement certains passages, notamment ceux où l'historien souhaite ralentir la narration et privilégier la description ou l'explicitation des causes d'un fait historique donné (comme, par exemple, dans le récit de la mort de Claude au livre 12 des *Annales*). En revanche, la séquence & +LN AD -LN, plus banale, se rencontre dans d'autres contextes, notamment lors de l'introduction de nouveaux personnages dans le récit, comme le montre, par exemple, l'examen du récit du début du règne de Néron (premiers chapitres du livre 13 des mêmes *Annales*). Sans l'aide de l'analyseur, il nous serait impossible, d'une part, de distinguer les deux structures et, d'autre part, de les dénombrer en vue de calculer les écarts existant sur ce point entre les textes de Tacite et ceux de ses prédécesseurs, chez qui ces mêmes structures peuvent également se rencontrer, mais avec une moindre fréquence et dans des proportions très différentes.

L'apport de l'analyseur *LatSynt* est donc primordial pour toutes les recherches visant à la segmentation et à la classification des textes latins. D'une manière plus globale, les recherches que nous menons montrent, nous semble-t-il, l'intérêt de ne pas cloisonner TAL et ADT : l'analyse des données textuelles est à la base même de la mise au point de l'analyseur, alors que celui-ci fournit en retour des informations indispensables pour pouvoir pratiquer une grammaticométrie véritable, prenant en compte à la fois la structure syntaxique et la topologie des textes.

Références

Bamman D. and Crane Gr. (2007). The Latin Dependency Treebank in a Cultural Heritage Digital Library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*. The Assocation for Computational Linguistics, pp. 33-40.

Bamman D., Passarotti M., Crane Gr. and Raynaud S. (2007). A Collaborative Model of Treebank Development. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT 2007)*, pp. 1-6.

Charpin F. (1977). *L'idée de phrase grammaticale et son expression en latin*. Paris : ANRT et Champion.

Charpin F. (1989). Étude de syntaxe énonciative : l'ordre des mots et la phrase. In Calboli, G., editor, *Subordination and Other Topics in Latin*. Amsterdam : John Benjamins, pp. 503-520.

Chausserie-Laprée J.P. (1969). *L'expression narrative chez les historiens latins, Histoire d'un style.* Paris : De Boccard.

Devine A.M. and Stephens L.D. (2006). *Latin Word Order: Structured Meaning and Information*. Oxford: Oxford University Press.

Longrée D. and Mellet S. (2009). Syntactical Motifs and Textual Structures. *Belgian Journal of Linguistics*, vol.(23) : 161-173.

Longrée D. and Mellet S. (2010a). Continuité et ruptures dans l'expression narrative des historiens latins. In Biraud, M., editor, *Phénomènes de continuité et de rupture : du lexique au discours. Études de linguistique latine et de linguistique grecque en hommage à Chantal Kircher*. Paris : L'Harmattan. À paraître.

Longrée D. and Mellet S. (2010b). Analysis of Textual Data, a interdisciplinary approach for studying the text structure indicators: the case of Latin historic narrative. In *Proceedings of the International Workshop on Multidisciplinary Approaches to Discourse 2010 (MAD'10): « Multidisciplinary Perspectives on Signalling Text Organisation »*, Moissac, March 17-20, 2010. À paraître.

Longrée D., Mellet S. and Luong X. (2008). Les motifs : un outil pour la caractérisation topologique des textes. In Heiden, S. et Pincemin, B., editors, *Actes des JADT 2008, 9èmes Journées internationales d'Analyse statistique des Données Textuelles*, Lyon, 12-14 mars, pp. 733-744.

Longrée D., Mellet S. and Poudat C. (2010). Les taggers, auxiliaires heuristiques en ADT. Dans ce volume.

Longrée D., Philippart de Foy C. and Purnelle G. (in press). Subordinate Clause Boundaries and Word Order in Latin: the Contribution of the L.A.S.L.A. Syntactic Parser Project LatSynt. In Anreiter, P. and Kienpointner, M., editors, *Proceedings of the 15th International Colloquium on Latin Linguistics*. Institut für Sprachen und Literaturen der Universität Innsbruck.

Longrée D., Philippart de Foy C. and Purnelle G. (en préparation). Prolepse et nature des subordonnants en Latin classique : l'apport du projet LatSynt. *Colloque international « Morphologie, syntaxe et sémantique des subordonnants »*, Clermont-Ferrand, 12-13 mars 2010.

Mazziotta N. (in press). Logiciel NotaBene pour l'annotation linguistique. Annotations et conceptualisations multiples. In Dumont, V. and Lejeune, C., editors, *Recherches qualitatives, Numéro Spécial : Logiciels pour l'analyse qualitative.*

Mellet S. and Longrée D. (2009). Syntactical Motifs and Textual Structures. Considerations based on the Study of a Latin historical Corpus. In Mellet, S. and Longrée, D., editors, *New Approaches in Text Linguistics = Belgian Journal of Linguistics / BJL 2009*: 161-174.

Panhuis D. (1982). *The Communicative Perspective in the Sentence, A Study of Latin Word Order*. Amsterdam: John Benjamins.

Spevak O. (2006). *L'ordre des constituants en latin, Aspects pragmatiques, sémantiques et syntaxiques*, HDR Université de Paris IV – Sorbonne.

Spevak O. (in press). *Constituent order in Classical Latin Prose*. Amsterdam: John Benjamins.