Les Études classiques 82 (2014), p. 47-63.

LE PROGRAMME « LATSYNT », UN INSTRUMENT POUR MESURER DES DISTANCES ENTRE PROSATEURS LATINS CLASSIQUES ?

Dès sa fondation en 1961, le LASLA ne se limita pas à la préparation d'index, mais exploita à d'autres fins les informations qui étaient alors encodées dans des fiches perforées. Ainsi, en appendice des deux premiers index publiés, consacrés aux *Consolations à Polybe* et *à Helvia* de Sénèque, trouvait-on des remarques sur la longueur des phrases dans ces deux œuvres. Dans un article publié un peu plus de vingt-cinq ans plus tard [1], Étienne Évrard, qui était à l'origine de ces remarques, en justifiait ainsi la présence :

> Il nous semblait, comme à de nombreux chercheurs, que la longueur des phrases était un caractère typique des textes littéraires et que l'étude devait en être fructueuse. Elle pouvait, à notre avis, se faire en deux directions diverses et, en un certain sens, complémentaires : tout d'abord, il est possible de déterminer la distribution globale des longueurs de phrases dans un texte et d'en tirer des indices tels que la longueur moyenne et la dispersion ; par ailleurs, on peut aussi s'intéresser à la manière dont les longueurs se succèdent effectivement dans le texte ; on a alors une étude pour ainsi dire chronologique susceptible d'être révélatrice à différents points de vue : la diversité éventuelle des parties d'une œuvre ; l'emploi de mini-patterns dans de petits groupes de phrases ; […] [2].

Comme ce fut souvent le cas, Ét. Évrard faisait là œuvre de pionnier en s'intéressant avant beaucoup d'autres à la linéarité du texte. Il s'agissait tout d'abord de caractériser un texte à la fois par la longueur de phrase calculée comme la moyenne des longueurs des phrases du texte et par l'ampleur des variations de longueur par rapport à cette moyenne. L'opération répétée sur divers textes devait permettre de les caractériser les uns par rapport aux autres et donc de les classer. Ensuite, l'objectif était d'étudier les variations

1. Cf. Ét. ÉVRARD, « L'étude des longueurs de phrases. Un réexamen des méthodes », *R.I.S.S.H*. 26 (1990), p. 55-66.

2. *Ibidem*, p. 55.

de longueur de phrases au fil du texte, à la fois pour segmenter l'œuvre en diverses parties en fonction de ces variations et pour rechercher des motifs transphrastiques caractéristiques. Ces objectifs restent encore tout-à-fait d'actualité dans les travaux actuels du LASLA.

Ét. Évrard était également conscient des limites d'une telle démarche dont il établit un inventaire dans le même article : d'une part, il n'est pas aisé de répartir des phrases en classes, car on peut hésiter sur les critères permettant de décider s'il s'agit d'une phrase courte ou d'une phrase longue, s'il s'agit d'une phrase isolée ou d'une phrase en séquence ; d'autre part, – et la question est plus cruciale –, la définition même de la notion de phrase ne va pas de soi en latin, et ce essentiellement pour deux raisons ; tout d'abord, comme l'a souligné Fr. Charpin[3], on ne trouve pas chez les grammairiens ou les rhéteurs anciens de concept correspondant exactement à ce que nous appelons aujourd'hui une phrase ; ensuite, on se heurte au côté parfois arbitraire de la ponctuation des éditeurs, en particulier quand il s'agit de savoir si un relatif postposé à la proposition principale introduit une subordonnée ou fonctionne comme un relatif de liaison introduisant une nouvelle indépendante et donc une nouvelle phrase.

Comme solution, Ét. Évrard préconisait une approche « dépendancielle » :

> La méthode que j'ai proposée consiste à doter chaque mot d'un descripteur supplémentaire qui n'est autre que le numéro d'ordre dans le texte du mot qui en est le régissant […] Dans la mesure où les numéros de régissants ont été correctement choisis et enregistrés, on a là un moyen de déterminer automatiquement, sur un critère rigoureux, les limites et la longueur des phrases.

En d'autres termes, Ét. Évrard avait inventé, bien avant tous, une forme de *Dependency Treebank*, mais comme tous les *Dependency Treebanks* actuels[4], le projet présentait un inconvénient majeur, celui de nécessiter un

3. Cf. notamment F. CHARPIN, « Étude de syntaxe énonciative : l'ordre des mots et la phrase », dans G. CALBOLI (éd.), *Subordination and Other Topics in Latin. Proceedings of the Third Colloquium on Latin Linguistics. Bologna*, *1-5 April 1985* (Studies in Language Companion Series, 17), Amsterdam - Philadelphie, John Benjamins, 1989, p. 503-520, en particulier p. 506-507.

4. Cf., entre autres, D. BAMMAN et Gr. CRANE, (2007). « The Latin Dependency Treebank in a Cultural Heritage Digital Library », dans C. SPORLEDER, A. VAN DEN BOSCH, Cl. GROVER (éd.), *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, Prague, Association for Computational Linguistics, 2007, p. 33-40 ; D. BAMMAN, M. PASSAROTTI, Gr. CRANE et S. RAYNAUD, « A Collaborative Model of Treebank Development», in *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT 2007)* [disponible en ligne à l'adresse suivante : http://tlt11.clul.ul.pt/ProceedingsTLT11.tgz].

encodage manuel important. Faute de moyens humains, le projet ne fut donc jamais mis en œuvre, mais l'idée maîtresse qui le sous-tendait, – étudier les structures syntaxiques de la phrase –, ne fut pas pour autant définitivement abandonné et finit par resurgir plus récemment sous une autre forme.

Depuis 2007, le LASLA s'est en effet lancé dans un vaste projet de recherche, celui de la mise au point d'un analyseur syntaxique automatique, baptisé « LatSynt »[5]. L'approche adoptée y est sensiblement différente : on s'appuie sur les informations morphosyntaxiques déjà encodées dans les fichiers du LASLA et sur la linéarité du texte pour tenter de préciser les structures syntaxiques des énoncés. Le projet s'inscrit dans une perspective non seulement de traitement automatique du langage (TAL), mais également d'analyse linguistique, puisqu'un de ses objectifs est d'éprouver la pertinence des diverses descriptions, publiées depuis les années quatre-vingt, qui tentent de rendre compte des règles régissant l'ordre des mots en latin. Cependant, dans la mesure où l'analyseur « LatSynt » vise à préciser des structures de phrases, on peut imaginer utiliser les nouvelles données produites par celui-ci pour caractériser, sur la base de nouveaux critères, divers types de textes. Dès lors, dans le cadre d'un volume consacré à la distance intertextuelle, il ne nous semblait pas inutile de tenter de préciser dans quelle mesure « LatSynt » pouvait se révéler un outil performant pour déceler des proximités ou des éloignements entre textes, en focalisant ici la recherche sur un corpus de prosateurs. Nous essayerons de répondre à cette question en deux étapes : dans un premier temps, nous présenterons le programme « LatSynt », en décrivant ses possibilités et ses limites au stade actuel de son développement ; nous évaluerons ensuite l'utilité que celui-ci peut avoir dès maintenant pour l'étude des distances entre textes.

5. D. LONGRÉE, C. PHILIPPART DE FOY et G. PURNELLE, « Structures phrastiques et analyse automatique des données morphosyntaxiques : le projet LatSynt », dans S. BOLASCO, I. CHIARI, L. GIULIANO (éd.), *Statistical Analysis of Textual Data. Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles, 9-11 June 2010, Sapienza University of Rome*, Rome, LED, 2010, p. 433-442 ; D. LONGRÉE, C. PHILIPPART DE FOY et G. PURNELLE, « Subordinate Clause Boundaries and Word Order in Latin: the Contribution of the L.A.S.L.A. Syntactic Parser Project LatSynt », dans P. ANREITER, M. KIENPOINTNER (éd.), *Proceedings of the 15th International Colloquium on Latin Linguistics* (Innsbrucker Beiträge zur Sprachwissenschaft, 137), Innsbruck, Institut für Sprachen und Literaturen der Universität Innsbruck, 2010, p. 673-681.

1. Objectifs, acquis et limites de la première phase du projet « LatSynt »

Conçu comme une recherche originale sur l'ordre des mots et sur les structures de l'énoncé latin, le projet « LatSynt » s'est donné plusieurs objectifs :

– développer des procédures d'analyse syntaxique automatisées proposant une alternative aux encodages manuels des projets de type *Treebank* ;

– évaluer la pertinence des descriptions linguistiques récentes ;

– fournir de nouveaux outils d'analyse des données textuelles (ADT), à la fois pour la modélisation des structures énonciatives et pour la classification et la segmentation des textes latins.

La première phase du projet a visé à délimiter les propositions subordonnées présentant un terme introducteur (conjonction de subordination, pronom, adjectif ou adverbe relatif ou interrogatif, particule interrogative) et à préciser leur niveau de subordination. La méthode utilisée pour ce faire s'appuie, d'une part, sur la linéarité du texte, d'autre part, sur les données alphanumériques – morphosyntaxiques et syntaxiques – de la base du LASLA. Dans les fichiers d'origine du LASLA, celles-ci se présentent sous la forme suivante :

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

On trouve ici six colonnes fournissant, pour chaque mot du texte :

(1) – le lemme, tel qu'il figure dans le dictionnaire choisi comme ouvrage de référence, à savoir le *Lexicon totius Latinitatis* de Forcellini (éd. de Corradini, Padoue, 1864) ;

(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, ou

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 ; en regard de chaque verbe subordonné, un code indique 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*).

Dans la figure précédente, &, dans la 6e colonne, signifie que *erant* est verbe de proposition principale ; le trait - signifie que la forme *possent* est un verbe subordonné ; LN indique que *possent* est subordonné par un pronom relatif *qui* ; le chiffre 32 après (*quibus*) indique le mode et le temps (3 subjonctif, 2 imparfait) du verbe que le relatif introduit. À partir de ces données, le programme « LatSynt » procède en plusieurs étapes au bornage automatique des propositions et à leur classement hiérarchique par l'attribution d'un niveau d'enchâssement syntaxique.

1.1. *Les étapes du bornage et du classement hiérarchique des propositions subordonnées*

Pour procéder à l'introduction d'un code de bornage, le programme « LatSynt » associe tout d'abord les informations syntaxiques du subordonnant et du verbe subordonné, en faisant figurer les mêmes données, par exemple pour *possent*, LN32, à la fois dans l'enregistrement du subordonné et dans celui du subordonnant. La deuxième étape consiste à produire un schéma linéaire de la structure syntaxique. Ainsi, par exemple, le passage suivant des *Annales* de Tacite (*Annales*, 13, 11, 2),

> […] *secutaque 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.*

sera réduit à la chaîne de codes :

> &0014 +LN14 -LN14 +LN12 +GK32 -GK32 -LN12

où & indique le verbe principal (*secuta*), où + indique un subordonnant (+LN14 = *quem* : +LN12 = *quas* ; +GK32 = *quam*), où - indique un verbe subordonné (-LN14 = *reddidit* : -LN12 = *uulgabat* ; +GK32 = *praeciperet*) et où 14 signale l'indicatif parfait, 12 l'indicatif imparfait et 32 le subjonctif imparfait.

La troisième étape vise à analyser les schémas obtenus et déduire les liens potentiels entre tout subordonnant et tout subordonné de même co-

dage. Sur la base de cette analyse, on peut ajouter au schéma des signes qui indiquent à la fois ces liens, les bornes théoriques des propositions et, le cas échéant, leur subordination relative (par inclusion des propositions les unes dans les autres). Pour le même exemple, on obtient ainsi la chaîne,

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

où les crochets obliques autour de <&0014> soulignent que *secuta (est)* correspond à la principale, où le crochet droit [devant +LN14 marque le début de la relative (LN14 indique toujours que *quem* introduit un verbe à l'indicatif parfait), et où le crochet droit] marque la fin de la proposition après le verbe *reddidit* à l'indicatif parfait. La séquence {+LN12 [+GK32 -GK32] -LN12} signale une subordonnée introduite par *quam* enchassée dans une relative.

La quatrième étape correspond au report de ce code-bornage dans la phrase proprement dite, ce qui donne le résultat suivant :

> <&*secuta* (*est*)> *que lenitas in Plautium Lateranum* [+*quem ob adulterium Messalinae ordine demotum -reddidit*] *senatui clementiam suam obstringens crebris orationibus* {+*quas Seneca testifi cando* [+*quam honesta -praeciperet*] *uel iactandi ingenii uoce principis -uulgabat*}

Ceci fait, en s'appuyant sur le code-bornage et sur la phrase appareillée, on peut produire un schéma indiquant le niveau de subordination de chaque proposition. Le schéma suivant propose une représentation d'un passage du *De Bello Gallico* (1, 12, 1).

<&0011>[+LN11-LN11]{+XK31[+YA31-YA31]=AG71-XK31}

0		*flumen <&est> Arar*
1	1	[+*quod per fines Haeduorum et Sequanorum in Rhodanum -influit*]
0		*incredibili lenitate ita*
1	2	{+*ut oculis in*
2	3	[+*utram partem -fluat*]
1	2	=*iudicari= non -possit*}

La première colonne indique le niveau de subordination, la deuxième, le numéro d'ordre de la subordonnée dans la phrase sur la base du subordonnant. Il reste alors au philologue à vérifier tout d'abord si les liens établis entre subordonnant et prédicat sont pertinents, ensuite si des éléments appartenant à la subordonnée ne se retrouvent pas soit à gauche du subordonnant, c'est-à-dire en dislocation à gauche (correspondant parfois à un accusatif proleptique), soit à droite du prédicat, c'est-à-dire en dislocation à droite ou postposition. Au stade actuel du développement de « LatSynt », ces deux derniers cas ne peuvent être pour l'instant repérés automatiquement et constituent une des limites majeures du bornage automatique.

1.2. *Les limites du bornage automatique*

S'appuyant sur le subordonnant et le prédicat subordonné, le bornage automatique ne permet pas de repérer les éléments disloqués à droite ou à gauche. Ces dislocations à gauche et à droite peuvent porter sur un syntagme nominal, comme c'est le cas dans l'exemple qui suit (*De Bello Gallico*, 6, 40, 2) où *castra*, le sujet de *sint*, n'est pas repéré comme faisant partie de la proposition en *quoniam* qui le précède.

[+XK31-XK31]<&0011>[+PX31-PX31][+CS34-CS34]<&0011>[+XK31-XK31-XK31]

0		*alii cuneo facto*
1	1	[+*ut celeriter -perrumpant*],
0		<&*censent*>
1	2	[+***quoniam*** *tam propinqua -sint*]
0		***castra*****,**
1	3	[+*etsi pars aliqua circumuenta -ceciderit*],
0		*at reliquos seruari posse* <&*confidunt*> *alii*
1	4	[+***ut*** *in iugo -consistant atque eundem omnes -ferant*]
0		***casum***

Si certains éléments comme *castra* peuvent être perçus comme « disloqués », c'est-à-dire n'occupant pas la place à laquelle on pourrait les attendre, d'autres peuvent ne pas être reconnus par « LatSynt » comme des éléments inclus dans la proposition subordonnée alors qu'ils occupent en postposition une position qui ne peut être perçue comme le fruit d'une dislocation. C'est le cas pour les propositions subordonnées dépendant elles-mêmes d'une subordonnée qui les précède et apparaissant donc comme des subordonnées postposées au verbe ou à la proposition dont elles dépendent. Ainsi dans l'exemple suivant (*Annales*, 13, 13, 1), la relative *ex cuius familiaribus* … n'est pas perçue comme faisant partie de la proposition subordonnée en *donec* qui précède :

&0071&0071[+JH12-JH12]&0071[+CA32-CA32-CA32][+LN15-LN15-LN15][+XK32{+LN12-LN12}-XK32]

0		*sed Agrippina libertam aemulam nurum ancillam alia que eundem in modum muliebriter* <&*fremere*> *neque paenitentiam filii aut satietatem* <&*opperiri*>
1	1	[+*quanto que foediora -exprobrabat*]
0		*acrius* <&*accendere*>
1	2	[+*donec ui amoris subactus -exueret obsequium in matrem se que Senecae -permitteret*]

0 *ex*
1 3 [+*cuius familiaribus Annaeus Serenus simulatione amoris aduersus eandem libertam primas adulescentis cupidines -uelauerat -praebuerat*]
0 *que nomen*
1 4 [+*ut*
2 5 [+*quae princeps furtim mulierculae -tribuebat*]
1 4 *ille palam -largiretur*]

Pour respecter la structure de la phrase, il faudrait pouvoir obtenir le schéma suivant :

&0071&0071[+JH12-JH12]&0071[+CA32-CA32-CA32{+LN15-LN15-LN15 [+XK32{+LN12-LN12}-XK32] }]

0 *sed Agrippina libertam aemulam nurum ancillam alia que eundem in modum muliebriter <&fremere> neque paenitentiam filii aut satietatem <&opperiri>*
1 1 [+*quanto que foediora -exprobrabat*]
0 *acrius <&accendere>*
1 2 [+*donec ui amoris subactus -exueret obsequium in matrem se que Senecae -permitteret*
0 *ex*
2 3 {+*cuius familiaribus Annaeus Serenus simulatione amoris aduersus eandem libertam primas adulescentis cupidines -uelauerat -praebuerat que nomen*
3 4 [+*ut*
4 5 {+*quae princeps furtim mulierculae -tribuebat*}
3 4 *ille palam -largiretur*] }]

Malheureusement, une telle complexité de phrase résiste encore actuellement au traitement automatique et le repérage des éléments postposés reste encore largement manuel. Pour un calcul de distances entre les textes, il n'est donc pas envisageable de s'appuyer ici sur la complexité de la structure des phrases et sur le nombre des niveaux de subordination rencontrés au sein de chaque phrase. En revanche, grâce à « LatSynt », nous disposons d'une autre donnée utile, à savoir la distance en nombre de mots entre le subordonnant et le dernier prédicat de la proposition. Nous pouvons également combiner ce critère avec celui de la longueur des phrases. Sur la base de ces deux critères, nous tenterons de calculer ici des distances intertextuelles dans deux corpus différents, en vue de tester la pertinence et l'efficacité de la méthode.

2. De nouveaux paramètres pour les calculs de distances intertextuelles

Nous avons tout d'abord choisi de tester ces paramètres sur un corpus composé uniquement de textes historiques, à savoir :

– César, *De Bello Gallico*, livres 1-7 (BG) ;
– Salluste, *Catilina - Iugurtha* (Cat - Jug) ;
– Quinte Curce, *Historia Alexandri Magni*, (QC) ;
– Tacite, *Agricola* (Agr) ;
– Tacite, *Annales*, livres 11-16 (Ann).

La méthode utilisée sera l'Analyse Factorielle des Correspondances appliquées à divers tableaux de distribution. La première figure présente la distribution, dans ces six ensembles textuels, de l'ensemble des propositions subordonnées classées en huit types en fonction de la distance entre le subordonnant et le dernier prédicat de la proposition : les distances entre subordonnant et dernier prédicat que nous avons retenues pour définir les huit classes sont de 0, 1, 2, 3, 4 mots, puis des 5 à 9, de 10 à 19 et de 20 à 89 mots.

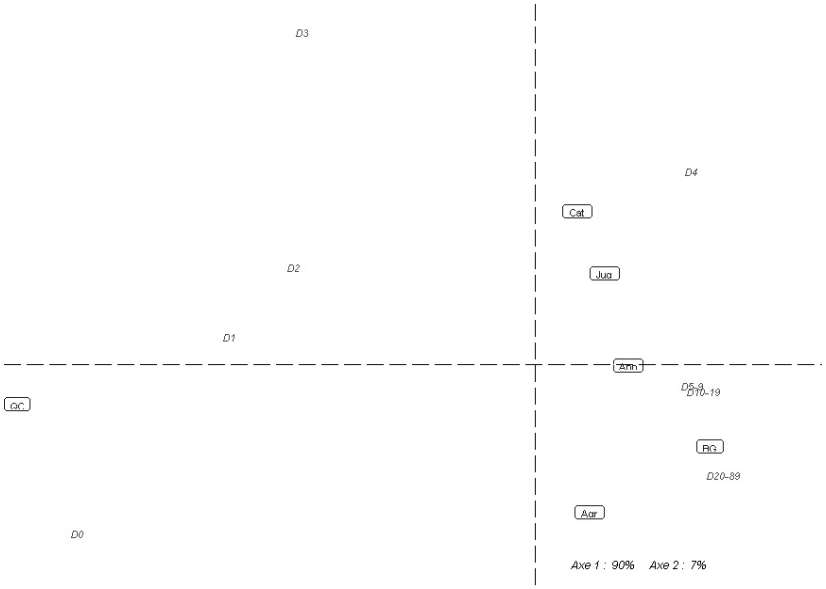

Figure 1 : AFC (axes 1 et 2) montrant la distribution, dans un corpus de six textes latins, des propositions subordonnées réparties en huit classes fixées par la distance subordonnant - prédicat

Le critère semble assez pertinent : Quinte-Curce, à gauche du graphe, utilise des propositions dans lesquelles le prédicat est proche du subordonnant, Salluste préfère des propositions où la distance subordonnant - dernier prédicat est de quelques mots, alors que César et Tacite privilégient des propositions où la distance entre le subordonnant et le dernier prédicat peut être très importante, jusqu'à quatre-vingt-neuf mots. Pour les propositions où le prédicat suit de peu le subordonnant, deux cas de figure peuvent se présenter : soit la proposition est elle-même courte, soit le verbe est antéposé au reste de la proposition (distance 0, éventuellement distance 1, notamment pour les prédicats constitués par une forme composée). Malheureusement « LatSynt » ne permet pas actuellement de distinguer automatiquement les deux cas.

Cette première approche peut être complétée par l'examen de la distance entre subordonnant et dernier prédicat pour les propositions subordonnées enchâssées dans une autre proposition subordonnée (c'est-à-dire en excluant les propositions de premier niveau de subordination et les propositions d'un niveau de subordination supérieur, mais qui ne sont pas enchâssées dans la proposition dont elles dépendent).

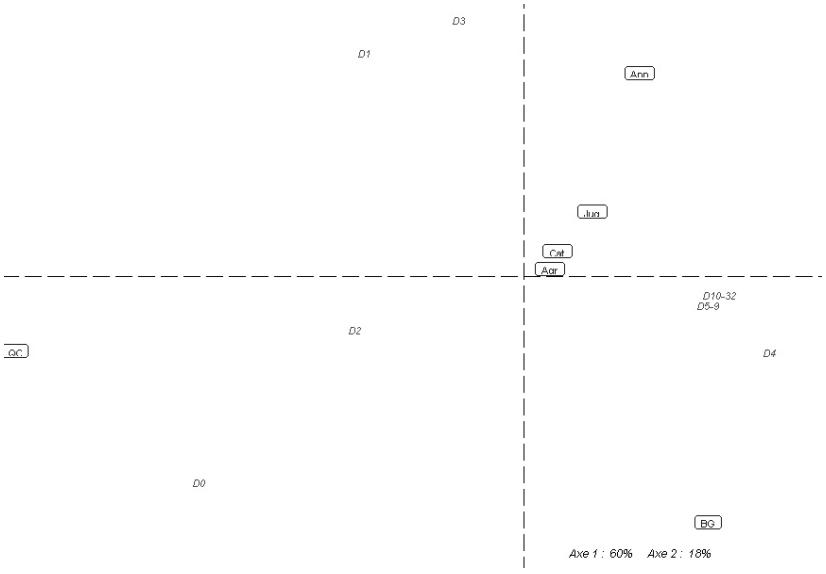

Figure 2 : AFC (axes 1 et 2) montrant la distribution, dans un corpus de six textes latins, des propositions subordonnées enchâssées dans une autre subordonnée, réparties en huit classes fixées par la distance subordonnant - prédicat

La distance maximale se réduit à trente-deux mots et la tendance générale est au raccourcissement de la distance, mais l'AFC obtenue montre que globalement les tendances restent les mêmes si ce n'est que les *Annales* s'écartent cette fois de la *Guerre des Gaules*.

Une autre manière d'affiner l'analyse est de focaliser celle-ci sur certains types de propositions. En raison de leur fréquence, nous avons choisi de nous intéresser ici aux propositions relatives et aux propositions en *cum*. En ce qui concerne les relatives, la figure suivante montre que sur l'axe 1, l'opposition reste la même entre propositions courtes à gauche avec Quinte-Curce et propositions plus longues à gauche avec l'ensemble des autres historiens. Toutefois, les pratiques de Salluste ne semblent plus ici être identiques ou proches dans ses deux œuvres, contrairement à ce qui se passait quand on considérait les propositions subordonnées dans leur ensemble : pour ce qui est des relatives, dans le *Catilina*, Salluste privilégie des propositions où le prédicat suit la conjonction à trois ou quatre mots de distance, alors que dans le *Iugurtha*, la distance peut-être beaucoup plus importante, ce qui rapproche ici le *Iugurtha* des *Annales* de Tacite. La chose n'a rien d'étonnant quand on sait que Salluste est considéré comme un des modèles de Tacite.

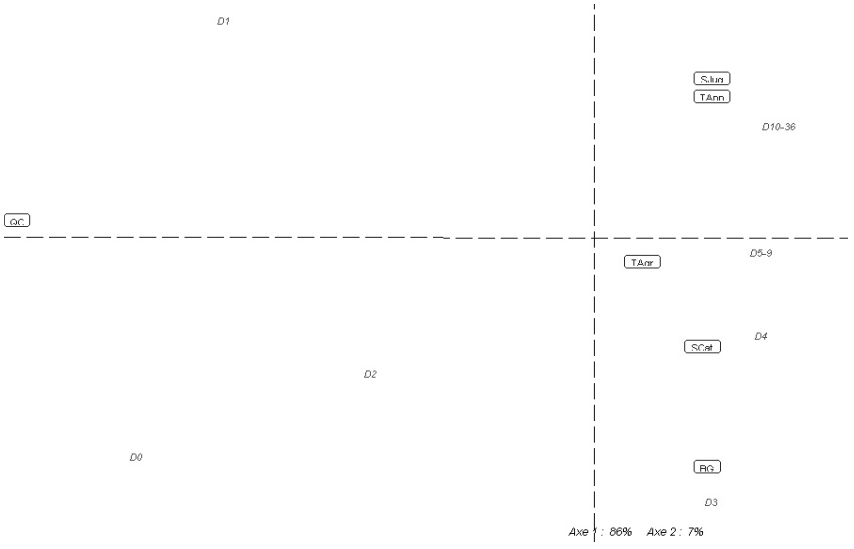

Figure 3 : AFC (axes 1 et 2) montrant la distribution, dans un corpus de six textes latins, des propositions relatives, réparties en huit classes fixées par la distance subordonnant - prédicat

Pour les propositions en *cum*, en revanche, les deux œuvres de Salluste, proches de l'origine des axes, semblent peu sensibles aux variations dans la distance entre le subordonnant et le prédicat. Pour le reste, Quinte-Curce semble toujours être l'auteur qui privilégie les distances les plus courtes au contraire de Tacite chez qui on rencontre, dans les *Annales*, les distances les plus longues. Un autre fait à noter est que, dans chacune des quatre AFC examinées jusqu'ici, l'*Agricola* se détache des *Annales*, ce qui peut s'expliquer sans doute tant par la différence générique entre les œuvres que par une évolution du style de leur auteur.

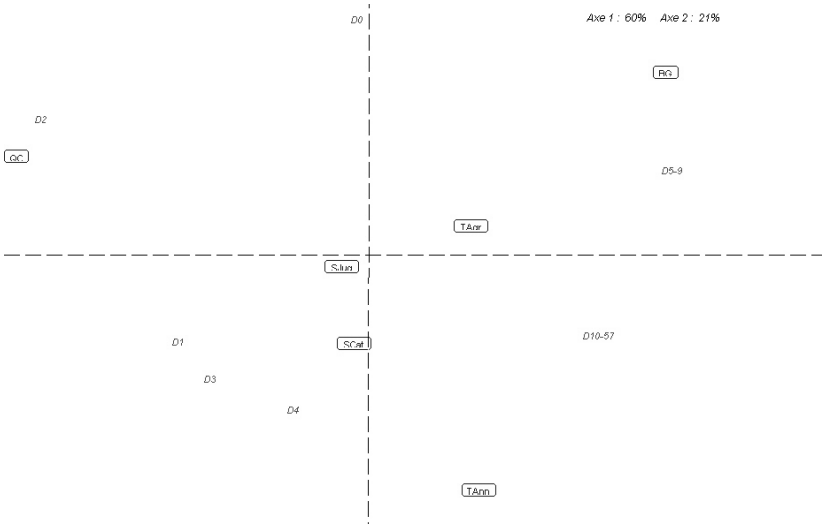

Figure 4 : AFC (axes 1 et 2) montrant la distribution, dans un corpus de six textes latins, des propositions en *cum*, réparties en huit classes fixées par la distance subordonnant - prédicat

Une question qui se pose ici est de savoir s'il existe une corrélation entre les distances relevées entre le subordonnant et le prédicat subordonné, et la longueur des phrases : un auteur qui écrit avec des subordonnées courtes choisit-il également d'écrire avec des phrases brèves ou multiplie-t-il les subordonnées courtes au sein de longues phrases ? L'AFC qui suit montre que Quinte-Curce semble bien préférer à la fois les phrases et les propositions courtes. En revanche, Tacite qui affectionne les subordonnées longues, n'emploie finalement que des phrases d'une longueur moyenne

comparées aux longues phrases de César, qui juxtapose de nombreuses subordonnées à la suite les unes des autres.

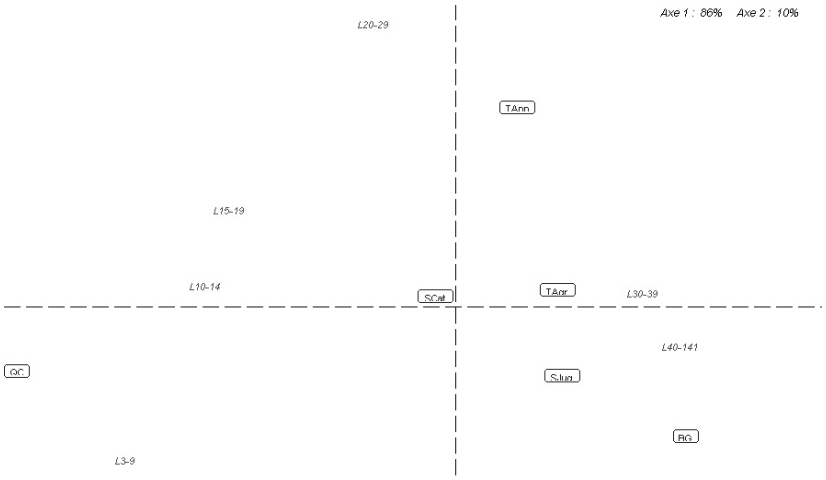

Figure 5 : AFC (axes 1 et 2) montrant la distribution, dans un corpus de six textes latins, des phrases réparties en six classes selon leur longueur.

*

* *

Sur un corpus d'historiens latins, la méthode semble donc pertinente et apporte un certain nombre d'informations sur les pratiques d'écriture des différents auteurs. On peut toutefois se demander si celle-ci peut être appliquée à un corpus plus hétérogène. Nous avons donc reproduit les tests précédents, mais cette fois sur un corpus élargi à quinze textes de prose, comportant également des traités et des discours :

– César, *Guerre des Gaules,* livres 1-7 (BG) ;
– Cicéron, *De amicitia* (Cam) ;
– Cicéron, *De senectute* (Csen) ;
– Cicéron, *Verrines* (CV1, CV21, CV22, CV23, CV24, CV25) ;
– Salluste*, Catilina - Iugurtha* (Cat - Jug) ;
– Quinte Curce*, Historia Alexandri Magni* (QC) ;
– Sénèque, *Epistulae* (Sepi) ;
– Tacite, *Agricola* (Agr) ;
– Tacite*, Annales*, livres 11-16 (Ann).

L'AFC suivante présente la distribution des textes sur la base de la distance subordonnant - prédicat pour toutes les propositions subordonnées. Sur l'axe 1, on constate une nette opposition entre les *Lettres à Lucilius*, à gauche, et presque toutes les autres œuvres à droite : plus encore que Quinte-Curce, Sénèque utilise dans ses *Lettres* des propositions où le prédicat apparaît à peu de distance du subordonnant. La proximité dans l'écriture des deux auteurs pourrait ici peut-être s'expliquer par leur proximité chronologique. Sur l'axe 2, la plupart des historiens s'opposent au reste de la prose. Les propositions subordonnées présentant la plus grande distance entre subordonnant et dernier prédicat se rencontrent essentiellement dans les discours de Cicéron.

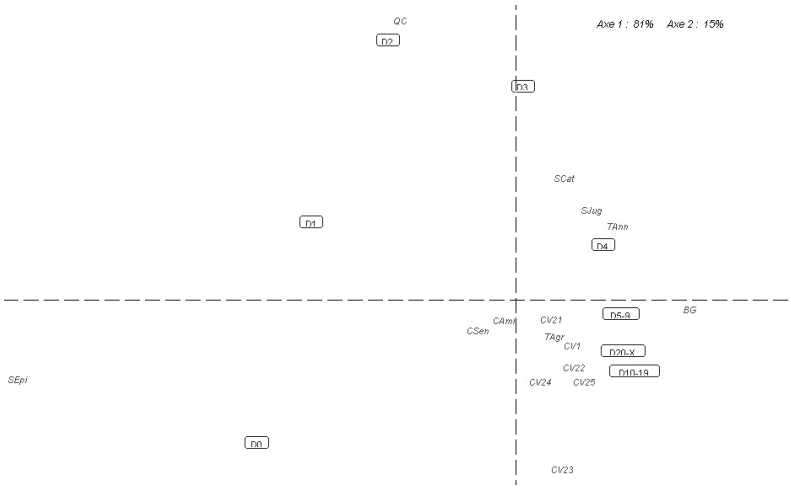

Figure 6 : AFC (axes 1 et 2) montrant la distribution, dans un corpus de quinze textes latins, des propositions subordonnées réparties en huit classes fixées par la distance subordonnant-prédicat

Ces résultats se confirment globalement quand on focalise l'analyse sur un type particulier de proposition, qu'il s'agisse des propositions relatives, comme dans la figure 7 ou des propositions en *cum* dans la figure 8. Dans le cas des relatives, les œuvres de Cicéron apparaissent mieux regroupées, à proximité des propositions relatives les plus longues. Dans le cas des propositions en *cum*, Quinte-Curce se rapproche des autres auteurs, laissant les *Lettres à Lucilius* tout à fait isolées à l'extrémité gauche de l'axe 1.

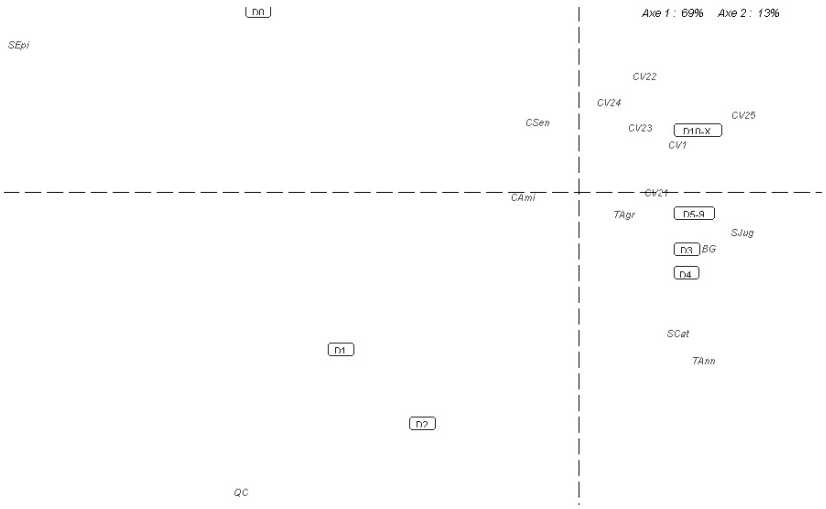

Figure 7 : AFC (axes 1 et 2) montrant la distribution, dans un corpus de quinze textes latins, des propositions relatives, réparties en sept classes fixées par la distance subordonnant-prédicat

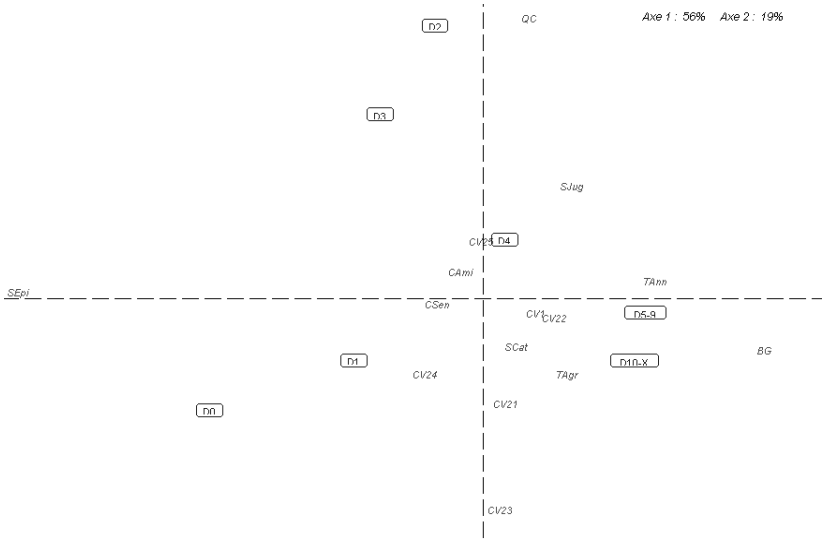

Figure 8 : AFC (axes 1 et 2) montrant la distribution, dans un corpus de quinze textes latins, des propositions en *cum*, réparties en sept classes fixées par la distance subordonnant - prédicat

Les figures 9 et 10 (réalisées avec le logiciel Anar, proposé par *Hyperbase*), présentant des graphes de distribution de sept classes de subordonnées en *cum* d'abord chez Sénèque, puis chez Quinte-Curce, montrent en effet que, sur ce point, les deux auteurs diffèrent : chez Sénèque, le prédicat suit le plus fréquemment immédiatement le subordonnant, alors que, chez Quinte-Curce, il suit souvent après un ou deux mots (sujet ou complément du verbe).

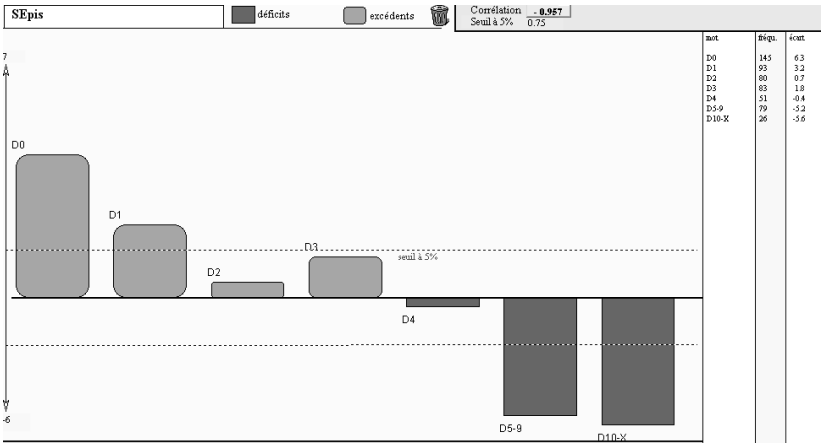

Figure 9 : Distribution, dans les *Lettres à Lucilius*, des propositions en *cum*, réparties en sept classes fixées par la distance subordonnant - prédicat

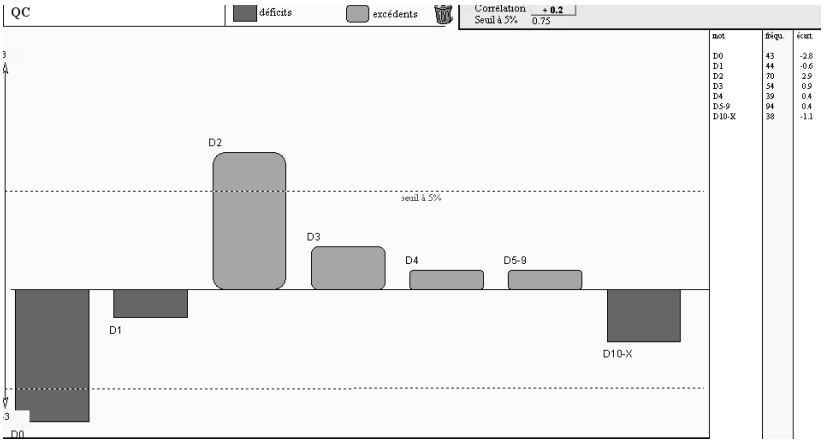

Figure 10 : Distribution chez Quinte-Curce des propositions en *cum*, réparties en sept classes fixées par la distance subordonnant - prédicat

3. Conclusion

La stabilité des résultats obtenus, tant en modifiant les paramètres que le corpus, montre que la méthode développée ici peut fournir un nouvel instrument de mesure des distances intertextuelles. Celle-ci permet de prendre en compte le texte dans sa linéarité, tant sur le plan micro-structurel de la position du prédicat par rapport au subordonnant que sur le plan plus large du rapport entre longueur de phrases et longueur des propositions. Elle fournit un critère de plus pour caractériser les différents styles des auteurs, mais peut également constituer un outil de différenciation générique.

Pour répondre aux souhaits formulés par Ét. Évrard et rappelés au début de cet article, les outils logiciels devront toutefois encore être affinés : il s'agirait de pouvoir établir avec précision le niveau de subordination de chaque proposition ainsi que sa longueur exacte, de manière notamment à distinguer les cas où le prédicat est antéposé ou placé dans les premiers mots d'une proposition qui pourrait par ailleurs être longue, de ceux où le prédicat clôt la proposition à quelques mots de distance à peine du subordonnant. Le travail doit être fait actuellement avec des procédures manuelles, mais on espère à moyen terme pouvoir automatiser, sinon totalement, du moins largement celui-ci. Sur le plan des analyses elles-mêmes, il faudrait non seulement élargir le corpus des textes traités, mais aussi s'intéresser de plus près aux variations entre les différents types de propositions : on pourra ainsi non seulement mieux préciser les distances entre textes selon l'époque, l'auteur ou le genre, mais également sans doute apporter de nouvelles informations utiles pour affiner la description de l'ordre des mots et du fonctionnement des subordonnées en latin.

Dominique LONGRÉE
Université de Liège, LASLA, 4000 Liège, Belgique
dominique.longree@ulg.ac.be
Gérald PURNELLE
Université de Liège, LASLA, 4000 Liège, Belgique
gerald.purnelle@ulg.ac.be