

Mesures de distance grammaticale entre les textes

Xuan Luong et Sylvie Mellet



Édition électronique

URL : <http://journals.openedition.org/corpus/34>

DOI : [10.4000/corpus.34](https://doi.org/10.4000/corpus.34)

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 15 décembre 2003

ISSN : 1638-9808

Référence électronique

Xuan Luong et Sylvie Mellet, « Mesures de distance grammaticale entre les textes », *Corpus* [En ligne], 2 | 2003, mis en ligne le 15 décembre 2004, consulté le 08 septembre 2020. URL : <http://journals.openedition.org/corpus/34> ; DOI : <https://doi.org/10.4000/corpus.34>

Ce document a été généré automatiquement le 8 septembre 2020.

© Tous droits réservés

Mesures de distance grammaticale entre les textes

Xuan Luong et Sylvie Mellet

- 1 Le calcul de distance entre les textes, développé à des fins de classification ou d'attribution d'auteur, est la démarche par laquelle on tente d'évaluer la plus ou moins grande ressemblance entre divers textes en prenant appui sur des éléments susceptibles d'être mesurés ou dénombrés, qui permettront d'une part de quantifier et ordonner ces degrés de ressemblance, d'autre part de reproduire le calcul autant de fois qu'on le souhaite sur différents types de textes en éliminant tout impact de la subjectivité.
- 2 Tout ou presque est dénombrable dans un texte : c'est le grand atout de la linguistique quantitative. En matière de distance entre les textes, la plupart des études se sont appuyées sur le lexique, compté en nombre de mots (N = total des occurrences constitutives de la chaîne linéaire du texte) ou en nombre de vocables (V = total des formes différentes constituant le vocabulaire spécifique du texte); le mot est le paramètre par excellence, facile à collecter et intuitivement pertinent : « No potential parameter of style below or above that of the word is equally effective in establishing objective comparison between authors and their common linguistic heritage »¹. D'autres travaux cependant ont pris appui sur des éléments infra-sémantiques tels que la longueur des phrases, la longueur des mots, le nombre de syntagmes prépositionnels, la proportion des mots-outils, la distribution des phonèmes à l'initiale des mots, etc.² Ces méthodes ont montré leur efficacité. Néanmoins, l'une et l'autre rencontrent aussi quelques limites : les distances lexicales sont extrêmement sensibles au thème et même au genre des textes comparés, si bien qu'en matière d'attribution d'auteur elles ne sont d'aucun secours si le texte en recherche de paternité et les textes de référence appartiennent à des genres différents³; elles ne sont pas très adaptées non plus lorsqu'on a affaire à des textes courts ou à des extraits. La seconde méthode est sans doute moins sensible aux phénomènes contextuels et donne des résultats plus stables à cet égard, mais elle défie toute tentative d'interprétation et ne permet que des classifications aveugles. Or, sans parler de la frustration éprouvée dans ce cas par le

linguiste, le littéraire ou l'historien, force est de constater que le calcul de distance a désormais des applications, telles que l'archivage automatique ou l'assistance à la distribution pertinente d'informations⁴, qui n'autorisent plus le recours exclusif à des critères de classification non interprétables.

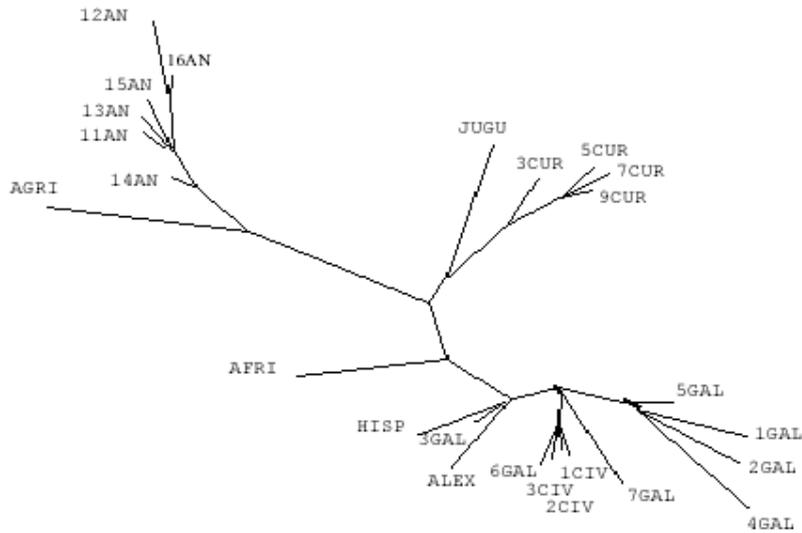
- 3 C'est pourquoi les catégories grammaticales et énonciatives, qui semblent davantage soustraites à l'influence du thème que le lexique et restent cependant porteuses de sens, pourraient offrir une troisième voie de recherche intéressante ; c'est à l'exploration de celle-ci que nous allons consacrer le présent article⁵, encouragés en outre par le constat de Cyril et Dominique Labbé, ici-même, que « les contributions, à la distance intertextuelle, des vocables classés en fonction de leurs catégories grammaticales sont assez souvent significativement différentes de la moyenne ».

1. Objectifs

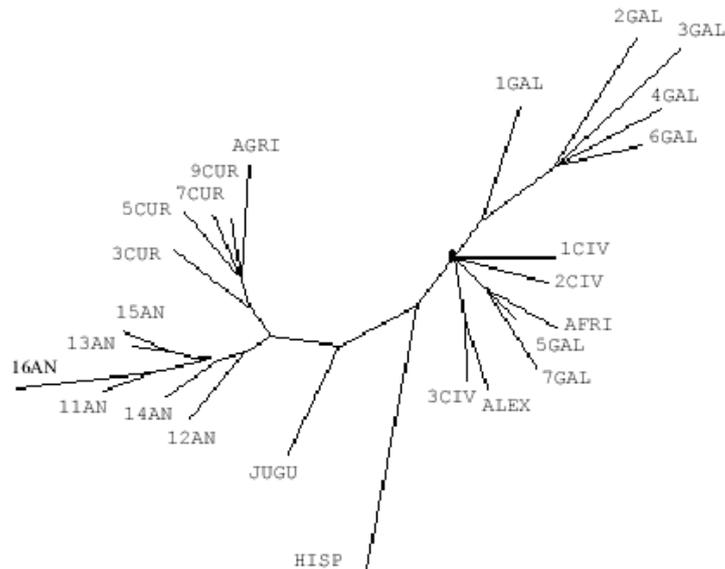
- 4 Il s'agira donc de voir s'il est possible de regrouper ou de différencier de manière pertinente des œuvres latines en fonction de la façon dont s'y distribuent certaines catégories grammaticales telles que les parties du discours, les types de subordonnées, les temps et les modes verbaux ou encore, puisque le latin est une langue flexionnelle, les cas nominaux.
- 5 Le choix de ces catégories est guidé par des considérations hétérogènes, il faut le reconnaître. Et ce, d'autant plus qu'on manque cruellement d'une définition théorique de ce que pourrait être le style idiosyncrasique d'un auteur en dehors du préformatage que lui imposent les diverses contraintes discursives et génériques⁶.
- 6 Le premier critère de choix sera donc purement contingent et dicté par des considérations pratiques : on ne retiendra ici que des éléments susceptibles d'être dénombrés automatiquement par un programme informatique de lecture du corpus⁷. Cette première condition nous offre cependant un éventail assez large dans la mesure où nous travaillons sur une base de textes lemmatisés et étiquetés, dans laquelle chaque mot est rapporté à son entrée de dictionnaire, indexé et suivi d'un code indiquant sa position dans le texte ainsi que les caractères essentiels de son analyse morpho-syntaxique⁸.
- 7 Le second critère sera l'existence d'analyses statistiques ou linguistiques préalables, en latin ou dans d'autres langues, ayant déjà suggéré le caractère distinctif de certaines oppositions grammaticales. Ainsi, É. Brunet a montré, à plusieurs reprises, sur des corpus littéraires français, l'importance de l'opposition entre un style nominal et un style verbal et, conclusion plus originale sans doute, le caractère très nettement distinctif, chez certains auteurs, de l'emploi du singulier et du pluriel. Les articles récents de D. Longrée et nos propres travaux sur l'emploi et la valeur des temps et modes verbaux en latin nous ont persuadée de leur poids discriminant ; ce constat est d'ailleurs corroboré, dans le domaine français, par certaines conclusions de la thèse de D. Mayaffre consacrée au discours politique français de l'entre-deux-guerres.
- 8 Nous avons donc finalement retenu les trois groupes de variables suivants dont la pertinence sera tour à tour évaluée : distribution des parties du discours⁹, distribution des cas et des nombres nominaux¹⁰, distribution des temps et modes verbaux combinés¹¹. Les structures de la langue imposent bien sûr des contraintes dans le choix de ces diverses formes ; mais il existe aussi une marge de liberté pour l'écrivain : en fonction du niveau de langue choisi, l'un emploiera plus volontiers l'ablatif – forme vivante et usuelle – là où un autre n'hésitera pas à recourir au locatif, forme archaïque et en voie de disparition ; en fonction de la structure narrative adoptée, laissant place

Figure 1

Analyse arborée des distances entre quelques textes d'historiens latins
en fonction de la distribution des parties du discours :
calcul sur les fréquences

**Figure 2**

Analyse arborée des distances entre quelques historiens latins en
fonction de la fréquence des cas et des nombres :
calcul sur les fréquences



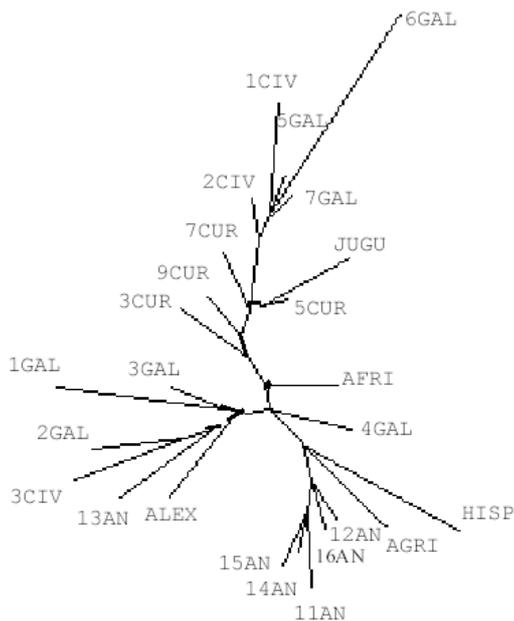
- 13 Les résultats sont apparemment très bons : la figure 1, en particulier, permet de regrouper à une extrémité de l'arbre tous les livres de César, avec une branche consacrée aux trois livres de la *Guerre Civile* auxquels s'adjoint le livre 6 de la *Guerre des Gaules* ; la *Guerre des Gaules* semble en effet plus hétérogène, mais ses sept livres restent néanmoins groupés au sein d'un même ensemble ; très proches et marquant pourtant leur différence, se trouvent aussi sur cette branche les trois ouvrages de même

inspiration et de même époque dus à des anonymes, lieutenants de César probablement (*Hisp.* ou *Guerre d'Espagne*, *Alex.* ou *Guerre d'Alexandrie* et *Afri.* ou *Guerre d'Afrique*). Cet ensemble césarien s'oppose très nettement à deux autres groupes, le plus proche étant constitué des quatre tomes de l'œuvre de Quinte-Curce, qui manifeste une certaine proximité avec la *Guerre de Jugurtha* de Salluste ; et beaucoup plus éloignés les textes de Tacite, avec une subdivision intéressante et littérairement pertinente entre la *Vie d'Agricola* d'une part (biographie laudative du beau-père de l'écrivain) et six livres des *Annales* d'autre part (récit année par année de l'histoire romaine depuis l'an 47 p.C. jusqu'en 66 p.C.). Les distances affichées ici correspondent donc tout à fait à la répartition du corpus selon les auteurs d'abord, selon les variations thématiques et de sous-genre ensuite.

- 14 La figure 2 affiche des résultats encore très pertinents, malgré quelques variations de détail : l'ensemble de l'œuvre de César et de ses épigones est toujours bien regroupé à une extrémité et s'oppose encore plus clairement à la fois à l'œuvre de Quinte-Curce et à celle de Tacite qui occupent chacune une branche à l'autre extrémité. La *Guerre de Jugurtha* de Salluste, plus isolée cette fois, se situe toujours à mi-chemin de ces deux pôles. La *Guerre d'Afrique*, excentrée dans la figure 1, s'agglomère parfaitement ici aux livres de César et c'est la *Guerre d'Espagne* qui s'en éloigne davantage. A l'autre bout de l'arbre, on observe une nouvelle configuration plus étonnante : la *Vie d'Agricola* de Tacite se trouve sur la branche de Quinte-Curce.
- 15 Ces petites variations sont sans doute le reflet de la pertinence linguistique ou stylistique des éléments de mesure choisis. Ceux-ci sont en effet de natures différentes et totalement indépendants les uns des autres. La convergence absolue des résultats n'est donc pas nécessaire et il n'y a pas lieu de l'exiger. Mais ces différences peuvent être dues également à un autre phénomène dont l'importance, encore mineure dans ce calcul, va devenir cruciale pour les tests suivants. On n'aura pas manqué d'observer en effet que, ne travaillant pas sur l'ensemble du vocabulaire des textes, mais sur certaines catégories grammaticales seulement, nous sommes amenés à calculer des distances non pas à partir de l'entier des textes, mais à partir de sous-ensembles constitués exclusivement des éléments textuels dans lesquels les catégories en question sont représentées¹⁴ : la taille du corpus s'en trouve réduite d'autant ; ainsi, si le premier calcul, opéré sur la distribution de l'ensemble des parties du discours, couvrait pratiquement la totalité des mots du corpus, en revanche le second, opéré sur les cas et nombres nominaux, ne s'intéressait qu'aux substantifs contenus dans le corpus. Dans le premier cas, le tableau de contingences totalisait 250 587 individus, dans le second 81 009 seulement. Cette diminution des effectifs va encore s'accroître dans les mesures portant sur les catégories verbales : le total du tableau de contingence chute à 57 203 individus pour les modes ou les voix, 44 307 pour les temps, 33 938 pour les personnes¹⁵ ; de l'analyse des parties du discours à celle des personnes verbales, c'est un peu comme si l'on traitait des textes huit fois moins longs.
- 16 Or c'est précisément une des limites de la méthode, soulignée par D. Labbé, que de perdre une partie de son efficacité sur les textes courts. On le vérifie en effet à la moindre structuration de l'arbre ci-dessous dans lequel les regroupements sont moins nets et moins pertinents que précédemment (livre 13 des *Annales* de Tacite au milieu du corpus césarien, par exemple, et *Guerre d'Espagne* au milieu de l'œuvre de Tacite).

Figure 3

Analyse arborée des distances entre quelques textes d'historiens latins
en fonction de la distribution des temps et modes verbaux :
calcul sur les fréquences



17 Nos textes latins ne sont pas très longs ; si on les réduit à des sous-ensembles constitués par les seuls mots porteurs de telle ou telle opposition grammaticale, on obtient des corpus de taille insuffisante pour pouvoir appliquer sans inquiétude le calcul de distance proposé par D. Labbé¹⁶. Bien sûr, on pourrait, par exemple, établir les fréquences relatives de chaque temps verbal par rapport à l'ensemble des formes constitutives de chaque texte (non seulement les verbes, mais les noms, les adjectifs, etc.) ; cette façon de procéder, qui aurait l'avantage de conserver la taille initiale des textes du corpus, a cependant un grave inconvénient : elle introduit subrepticement un paramètre parasite, à savoir – dans l'exemple choisi – le caractère plus ou moins verbal du style de l'auteur ou, en d'autres termes, la place plus ou moins grande occupée dans les textes par les formes verbales au regard des formes nominales, pronominales ou autres.

18 C'est pourquoi nous avons cherché une autre méthode d'analyse en partant de l'examen renouvelé des tableaux de contingence initiaux.

3. Propositions et tests d'évaluation 3.1. Exploitation des tableaux de contingence

19 Considérons les données du corpus césarien¹⁷ ; nous disposons de relevés automatiques d'effectifs que récapitulent les tableaux suivants : pour chaque texte en ligne les colonnes présentent le nombre absolu d'occurrences de nominatif singulier, nominatif pluriel, etc. ou d'indicatif présent, indicatif imparfait, etc. Comme on le constate aisément, ces tableaux récapitulatifs offrent des colonnes bien fournies et quelques colonnes aux effectifs très faibles – et ce, malgré les regroupements effectués sur ces colonnes creuses (regroupement du singulier et du pluriel pour le vocatif et pour le locatif, regroupement du futur et du futur antérieur sous l'étiquette 'IndFut', de l'impératif présent et de l'impératif futur sous l'étiquette 'Impér.').

Effectifs absolus des cas et nombres nominaux dans le corpus césarien

| | Nom. Sg. | Nom. Pl. | Voc. | Acc. Sg. | Acc. Pl. | Gén. Sg. | Gén. Pl. | Dat. Sg. | Dat. Pl. | Abl. Sg. | Abl. Pl. | Loc |
|------|-------------|-------------|------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-----|
| AFRI | 396 | 200 | 11 | 829 | 539 | 329 | 235 | 117 | 56 | 1042 | 495 | 11 |
| ALEX | 387 | 139 | 1 | 737 | 312 | 314 | 172 | 105 | 53 | 748 | 384 | 3 |
| 1CIV | 400 | 195 | 0 | 706 | 424 | 365 | 204 | 72 | 48 | 746 | 444 | 8 |
| 2CIV | 216 | 121 | 2 | 399 | 212 | 210 | 127 | 50 | 37 | 502 | 219 | 6 |
| 3CIV | 560 | 210 | 2 | 987 | 520 | 480 | 235 | 128 | 75 | 1127 | 592 | 32 |
| 1GAL | 216 | 125 | 0 | 557 | 322 | 217 | 147 | 70 | 60 | 526 | 242 | 3 |
| 2GAL | 106 | 69 | 0 | 251 | 220 | 101 | 81 | 25 | 29 | 284 | 166 | 0 |
| 3GAL | 120 | 99 | 0 | 185 | 154 | 101 | 83 | 25 | 23 | 232 | 168 | 0 |
| 4GAL | 121 | 77 | 1 | 255 | 185 | 99 | 96 | 32 | 22 | 249 | 201 | 2 |
| 5GAL | 240 | 111 | 1 | 445 | 268 | 194 | 146 | 73 | 32 | 556 | 271 | 1 |
| 6GAL | 165 | 104 | 1 | 317 | 244 | 156 | 149 | 38 | 44 | 373 | 216 | 1 |
| 7GAL | 341 | 209 | 1 | 702 | 427 | 333 | 219 | 94 | 112 | 932 | 467 | 15 |
| HISP | 223 | 136 | 0 | 430 | 201 | 112 | 105 | 33 | 17 | 508 | 146 | 0 |

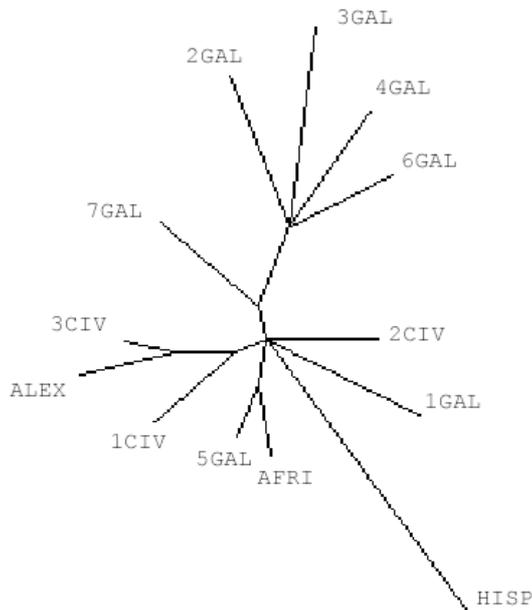
Effectifs absolus des temps et modes verbaux dans le corpus césarien

| | Ind. Prt. | Ind. Impf. | Ind. Parf. | Ind. Ppn. | Ind. Fut. | Subj. Prt. | Subj. Impf. | Subj. Parf. | Subj. Ppn. | Inf. Prt. | Inf. Parf. | Inf. Fut. | Impér. |
|------|-----------|------------|------------|-----------|-----------|------------|-------------|-------------|------------|-----------|------------|-----------|--------|
| AFRI | 365 | 243 | 333 | 143 | 7 | 22 | 279 | 5 | 91 | 438 | 43 | 24 | 5 |
| ALEX | 279 | 302 | 292 | 96 | 2 | 14 | 322 | 5 | 91 | 278 | 54 | 33 | 1 |
| 1CIV | 549 | 298 | 128 | 132 | 0 | 134 | 194 | 40 | 51 | 388 | 68 | 39 | 0 |
| 2CIV | 295 | 151 | 146 | 68 | 0 | 58 | 127 | 12 | 33 | 196 | 26 | 14 | 2 |
| 3CIV | 205 | 440 | 613 | 209 | 7 | 15 | 444 | 3 | 107 | 454 | 61 | 41 | 9 |
| 1GAL | 204 | 142 | 214 | 64 | 0 | 72 | 287 | 31 | 101 | 357 | 76 | 51 | 0 |
| 2GAL | 54 | 96 | 136 | 51 | 0 | 14 | 128 | 11 | 27 | 178 | 24 | 13 | 0 |
| 3GAL | 87 | 99 | 103 | 35 | 0 | 25 | 92 | 3 | 29 | 134 | 19 | 5 | 0 |
| 4GAL | 118 | 102 | 180 | 66 | 1 | 37 | 115 | 8 | 41 | 170 | 18 | 18 | 1 |
| 5GAL | 357 | 134 | 164 | 81 | 4 | 83 | 156 | 17 | 47 | 284 | 44 | 27 | 1 |
| 6GAL | 406 | 78 | 103 | 53 | 0 | 75 | 73 | 12 | 16 | 176 | 40 | 15 | 2 |
| 7GAL | 522 | 245 | 236 | 131 | 3 | 142 | 254 | 26 | 67 | 407 | 65 | 37 | 8 |
| HISP | 172 | 124 | 298 | 28 | 6 | 8 | 146 | 4 | 119 | 168 | 37 | 4 | 0 |

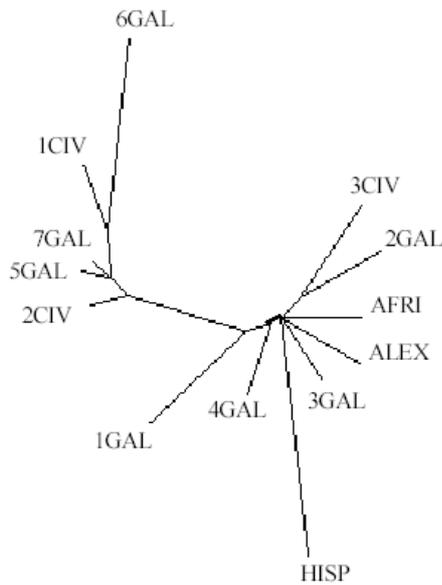
- 20 En l'absence de ces colonnes à faibles effectifs, ce type de données se prêterait assez bien à un simple calcul de distance par le χ^2 . Et, de fait, les résultats obtenus par ce calcul classique, en supprimant les colonnes vides, ne sont pas inintéressants ; voici la représentation arborée des deux tableaux ci-dessus :

Figure 4

Analyse arborée des distances au sein du corpus césarien en fonction de la distribution des cas et nombres : calcul du χ^2 sur le tableau de contingence (excepté les vocatifs et locatifs)

**Figure 5**

Analyse arborée des distances au sein du corpus césarien en fonction de la distribution des temps et modes du verbe : calcul du χ^2 sur le tableau de contingence (excepté les futurs et l'impératif)



- 21 Sur ces figures 4 et 5, le point intéressant est la tension entre stabilité et variation. Cette dernière est due, bien sûr, au choix de critères d'analyse différents qui ne mettent donc pas en évidence exactement les mêmes proximités ou éloignements. Mais derrière cette variation attendue, on discerne des lignes de force stables : la grande proximité de la *Guerre d'Alexandrie* et de la *Guerre d'Afrique* qui se mélangent assez aisément avec certains livres de César lui-même alors que la *Guerre d'Espagne* se situe toujours à une

distance remarquable de l'ensemble : les trois pseudo-César n'avaient donc pas la même qualité d'imitation. On notera aussi la proximité récurrente des trois livres centraux de la *Guerre des Gaules* (2GAL, 3GAL, 4GAL) et la relative dispersion des trois livres de la *Guerre Civile*, confirmant ainsi les conclusions auxquelles avaient depuis longtemps abouti les philologues¹⁸.

- 22 Cependant, malgré l'intérêt de telles analyses, le recours au simple calcul du χ^2 n'est pas vraiment satisfaisant. En effet, il oblige, comme on l'a dit, à laisser de côté des colonnes qui, certes, rassemblent des effectifs très faibles, mais contiennent des informations importantes aux yeux du philologue et du stylisticien. Consultons à nouveau les deux tableaux de contingence ci-dessus : il n'est vraiment pas anodin pour l'analyse linguistique de constater que les épigones de César emploient à eux trois davantage de futurs de l'indicatif que ne le fait César dans l'ensemble de ses dix livres ; que l'auteur de la *Guerre d'Afrique* utilise volontiers le vocatif alors que celui de la *Guerre d'Espagne* l'évite. Il faudrait donc pouvoir récupérer ces informations dans le calcul de distance, tout en sachant que les effectifs seront toujours trop bas pour permettre les traitements statistiques classiques.
- 23 C'est pourquoi nous avons pensé faire porter le raisonnement non pas sur les valeurs chiffrées, mais sur leur relation d'ordre. Il s'agira donc de combiner l'examen traditionnel du profil de chaque ligne avec un classement opéré selon leur rang vis à vis de l'emploi de telle ou telle catégorie grammaticale.

3.2. Calcul de distance à partir d'un classement ordinal

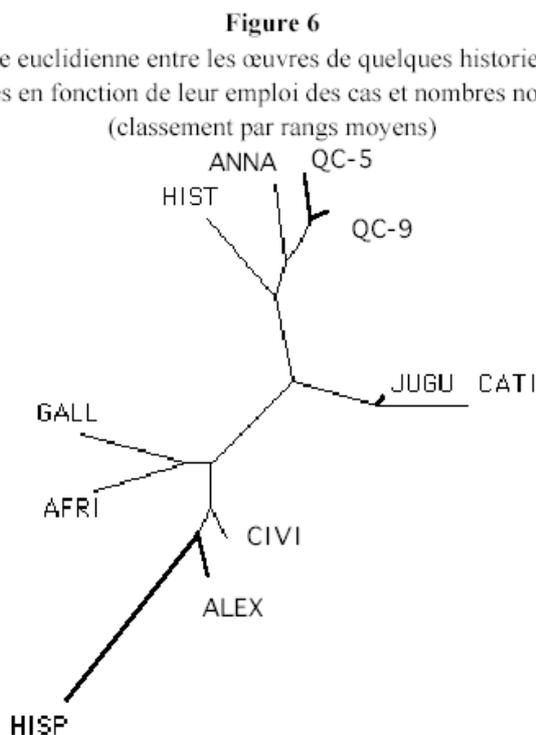
- 24 A partir des fréquences relatives de chacune des catégories rassemblées au sein d'un même tableau, on peut affecter à chaque texte un numéro d'ordre sur chaque colonne constitutive de ce tableau. On assimile ainsi les catégories grammaticales à des « juges » et les textes à des « candidats » classés par les « juges » : est classé premier le texte qui, proportionnellement, recourt le plus à la catégorie désignée et dernier celui qui l'emploie le moins souvent. Ce classement peut aboutir d'abord à un *préordre* s'il donne lieu à des ex-aequo ; il peut être ensuite transformé en classement par « rangs moyens » (la série 1 - 2 - 3 - 3 - 5 - 6 - 6 - 6 - 9 - 10 - 11 devenant alors 1 - 2 - 3,5 - 3,5 - 5 - 7 - 7 - 7 - 9 - 10 - 11). Dans tous les cas, on travaillera sur des données parfaitement homogènes et distribuées sur une échelle réduite : on n'aura donc guère motif à pondérer ces données et un simple calcul de distance euclidienne, par exemple, sera parfaitement légitime.
- 25 Une objection vient toutefois à l'esprit : ce classement ordinal, projetant les textes sur une échelle régulière dont les degrés sont équidistants, fait subir aux données brutes une sorte de lissage qui atténue les écarts susceptibles de séparer les textes. Pour prendre un exemple, quatre textes pour lesquels la catégorie C1 aurait respectivement une fréquence d'emploi de 18%, 18,5%, 18,4% et 17,8% des occurrences et pour lesquels la catégorie C2 aurait respectivement la fréquence 6,5%, 18,5%, 18,4% et 6% obtiendraient exactement le même classement ordinal dans les deux colonnes, à savoir : 3 - 1 - 2 - 4. Pour cette raison on a également envisagé et testé la possibilité d'un classement « subjectif » qui donne à l'analyste la liberté de regrouper les textes selon des classes de fréquence ; dans le cas d'école donné ci-devant en exemple, pour la catégorie C1 les quatre textes appartiendraient à la même classe et seraient tous classés ex-aequo ; pour la catégorie C2 en revanche, le premier et le dernier texte appartiendraient évidemment à la même classe, celle des textes à faible effectif, et les

deux textes centraux à une autre classe, à plus fort effectif. Le préordre deviendrait alors : 3 - 1 - 1 - 3 et le classement par rangs moyens : 3,5 - 1,5 - 1,5 - 3,5.

- 26 Notre objet est toujours ici de construire des distances et des similarités entre les textes ainsi classés ; il convient d'observer cependant que la méthode permettrait aussi de mettre en évidence les proximités et les éloignements entre les catégories.

3.3. Test et évaluation sur les historiens latins

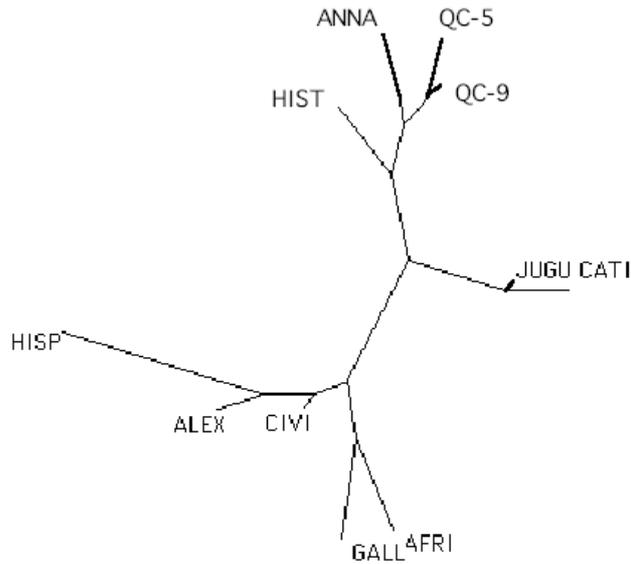
- 27 3.3.1. Testons d'abord notre méthode sur un corpus regroupant les œuvres de différents historiens latins : le genre sera donc le même (narration historique en prose), mais le sous-genre sera différent (Annales, commentaires, biographies), les auteurs et les dates également. Les oppositions devraient donc être marquées, l'arbre clairement structuré : nous éviterons ainsi que des incertitudes philologiques viennent perturber l'évaluation de l'outil. La catégorie grammaticale retenue est celle des cas et nombres nominaux.



- 28 Les résultats sont pleinement satisfaisants : comme précédemment, les œuvres de César et de ses épigones s'opposent à celles de Tacite (HIST et ANNA) et de Quinte-Curce (QC-5 et QC-9), Salluste (JUGU et CATI) se situant à mi-chemin. Au sein du groupe césarien, la *Guerre des Gaules* et la *Guerre Civile* se différencient nettement, la première étant flanquée de la *Guerre d'Afrique*, la seconde de la *Guerre d'Alexandrie* et la *Guerre d'Espagne* se situant assez loin, en marge du groupe, tout cela conformément à nos attentes¹⁹.
- 29 Il est à noter que le calcul de la distance euclidienne sur un classement « subjectif » (voir plus haut) donne exactement la même figure en (7) qu'en (6), prouvant à la fois que ce classement est légitime, mais aussi qu'il n'apporte pas, ici du moins, d'amélioration majeure à la perception des distances.

Figure 7

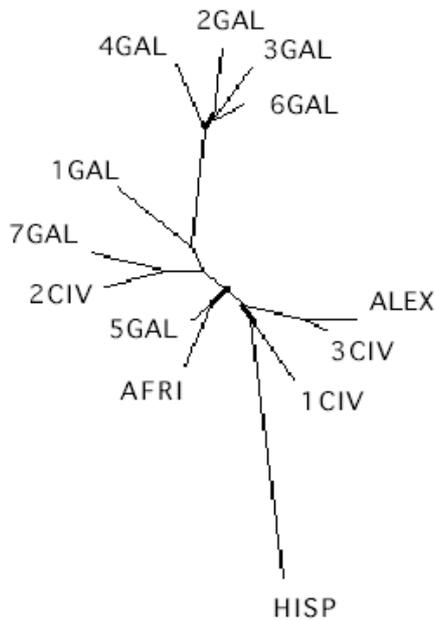
Distance euclidienne entre les œuvres de quelques historiens latins classées en fonction de leur emploi des cas et nombres nominaux (classement « subjectif » par classes de fréquence)



- 30 3.3.2. Voyons maintenant si notre outil est capable de rendre compte d'oppositions ou de proximités plus subtiles, c'est-à-dire s'il peut nous permettre de visualiser la complexe hétérogénéité des commentaires césariens. Le paramètre de classement sera toujours le même, à savoir les catégories nominales.

Figure 8

Distance euclidienne entre les œuvres de César et de ses épigones classées en fonction de leur emploi des cas et nombres nominaux (classement par rangs moyens)



- 31 Il convient de comparer cette figure avec la figure 4 élaborée à partir des mêmes unités d'analyse : on retrouve le même groupement fortement structuré des livres 2, 3, 4 et 6

de la *Guerre des Gaules*, dont les distances se sont raccourcies ; la même proximité du livre 5 de la *Guerre des Gaules* et de la *Guerre d'Afrique*, puis des livres 1 et 3 de la *Guerre Civile* auxquels s'adjoint la *Guerre d'Alexandrie* ; le même éloignement marqué de la *Guerre d'Espagne*. En revanche, le livre 2 de la *Guerre Civile* se trouve désormais rattaché à la branche du livre 7 de la *Guerre des Gaules* au lieu d'être autonome dans une structure étoilée comme précédemment : cette méthode, associée à la prise en compte des occurrences de vocatifs et de locatifs, semble donner une structuration plus ferme de l'arbre avec des proximités mieux affichées.

- 32 Qu'en est-il lorsque le classement est effectué sur les catégories verbales ? Comparons la figure 9 à la figure 5 :
- 33 Les différences sont un peu plus marquées ; cette fois-ci, contrairement au cas précédent, les formes rares introduisent plus de variations et des écarts plus grands entre les textes. Le groupe qui associe les trois derniers livres de la *Guerre des Gaules* et les deux premiers livres de la *Guerre Civile* est identique, quoique moins fortement structuré. En revanche, l'autre partie de l'arbre fait apparaître ici des proximités qui se devinaient mal dans la figure 5 : compagnonnage dans l'excentricité de la *Guerre d'Espagne* et de la *Guerre d'Alexandrie*, alors que la *Guerre d'Afrique*, effectivement réputée pour être mieux écrite à l'imitation de César, reste toujours rattachée à la branche qui porte aussi le livre 4 de la *Guerre des Gaules* et le livre 3 de la *Guerre Civile*. Enfin les trois premiers livres de la *Guerre des Gaules* sont aussi mieux regroupés.

Conclusion

- 34 Nous avons donc tenté d'élaborer un calcul de la distance grammaticale entre les textes. Cet outil, soyons honnêtes, n'a pas de performances exceptionnelles, notamment en matière d'attribution d'auteur²⁰. Comme tous ceux dont on dispose aujourd'hui, son pouvoir discriminant n'est pas assez fort pour reconnaître et isoler nettement un imitateur habile²¹ au sein d'un corpus qui ou bien est dominé par les oppositions d'époques et de sous-genres (fig. 7), ou bien, en raison d'un zoom excessif, trahit principalement l'hétérogénéité globale d'une œuvre (fig. 9).
- 35 Néanmoins, notre méthode par classement ordinal fournit des représentations de distances interprétables et remarquablement stables sur les différents corpus étudiés. Par ailleurs, en matière de classement générique et/ou chronologique, cette méthode obtient des résultats largement aussi bons que les autres méthodes connues : les œuvres sont toujours regroupées selon leurs auteurs dans une configuration qui oppose la période classique à la période impériale ; et la marginalité de la *Guerre d'Espagne* dans le corpus césarien est toujours clairement mise en évidence (figures 6, 7 et 8). On peut donc affirmer que cette méthode par classement ordinal permet de réaliser un calcul fiable sur des données numériques globalement peu importantes et sur des paramètres très inégalement répartis à travers les textes – deux caractéristiques auxquelles est souvent confronté le philologue, mais qui perturbent dangereusement les autres calculs de distance. Elle offre donc une alternative intéressante, notamment lorsque, pour une raison ou une autre, on souhaite échapper à l'hégémonie de la distance lexicale.

BIBLIOGRAPHIE

- Baayen H., Van Halteren H. & Tweedie F. (1996). « Outside the Cave of Shadows. Using Syntactic Annotation to Enhance Authorship Attribution », *Literary and Linguistic Computing* 11 : 121-131.
- Baayen H., Van Halteren H., Neijt A. & Tweedie F. (2002). « An Experiment in Authorship Attribution », in A. Morin & P. Sébillot (éds.), *JADT 2002, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*. Saint-Malo : Irista, Inria, vol. 1 : 69-75.
- Barthélémy J.-P. & Luong X. (1987). « Sur la topologie d'un arbre phylogénétique : aspects théoriques, algorithmes et applications à l'analyse des données textuelles », *Mathématiques et Sciences humaines* 100 : 57-80.
- Beauvisage T. (2001). « Exploiter des données morpho-syntaxiques pour l'étude statistique des genres. Application au roman policier », *TAL* 42, 2 : 579-608.
- Brunet É. (1988a). « Une mesure de la distance intertextuelle : la connexion lexicale », *Revue Informatique et Statistique dans les Sciences humaines (Le nombre et le texte. Hommage à Étienne Évrard)* 24, 1-4 : 81-116.
- Brunet É. (2002). « Un texte sacré peut-il changer ? Variations sur l'Évangile », in J. Cook (éd.), *Bible and Computer*. Leiden / Boston : Brill, pp. 79-98.
- Brunet É. & Muller C. (1988). « La statistique résout-elle les problèmes d'attribution ? », *Strumenti critici* III, 3 : 367-387.
- Burrows J. (2003). « Questions of authorship: attribution and beyond », *Computers and the Humanities* 37, 1 : 5-32.
- Dubrocqard M., Luong X. & Cottier J.-F. (2002). « Statistique et attribution de textes : l'analyse des formes peut-elle remplacer celle des lemmes ? Le cas des textes attribués à Raoul le Moine (XII^{ème} siècle) », *Médiévales* 42 (*Le latin dans les textes*) : 55-71.
- Holmes D. (1998). « The Evolution of Stylometry in Humanities Scholarship », *Literary & Linguistic Computing* 13, 3 : 111-118.
- Labbé D. & Monière D. (2000). « La connexion intertextuelle. Application au discours gouvernemental québécois », in M. Rajman & J.-C. Chappelier (éds.), *JADT 2000. Actes des 5èmes journées internationales d'Analyse statistique des Données Textuelles*. Lausanne : EPFL, vol. 1 : 85-94.
- Literary & Linguistic Computing* 13, 3 (1998).
- Longrée D. (2003, à paraître). « Spécificités stylistiques et distributions temporelles chez les historiens latins : sur les méthodes d'analyse quantitative d'un corpus lemmatisé », *Actes des 2^{èmes} Journées de linguistique de corpus* (Lorient, 12-14 sept. 2002). Rennes : Presses Universitaires de Rennes.
- Luong X. (éd.) (1989). *Analyse arborée des données textuelles, Tree Analysis of Textual Data*. CUMFID 16, Nice.
- Luong X. (1994). « L'analyse des données textuelles : mode d'emploi », *Travaux du Cercle linguistique de Nice* 16 : 25-42.
- Malrieu D. & Rastier F. (2001). « Genres et variations morpho-syntaxiques », *TAL* 42, 2 : 547-577.

- Mellet S. (1996). « Les atouts de la lemmatisation », in G. Moracchini (éd.) *Actes du Colloque international « Bases de données linguistiques : conceptions, réalisations, exploitations »*. Univ. de Corse / Univ. de Nice - Sophia Antipolis, pp. 309-316.
- Mellet S. (2002). « La lemmatisation et l'encodage grammatical permettent-ils de reconnaître l'auteur d'un texte ? », *Médiévales* 42 (*Le latin dans les textes*) : 13-26.
- Rastier F. (2001). « Vers une linguistique des styles », *L'Information Grammaticale* 89 : 3-6.
- Somers H. & Tweedie F. (2003). « Authorship Attribution and Pastiche », *Computers and the Humanities* 37, 4 : 407-429.
- Stamatos E., Fakotakis N. & Kokkinakis G. (2001). « Computer-based Authorship Attribution without lexical Measures », *Computers and the Humanities* 35, 2 : 193-214.

NOTES

- 1.. D.R. Tallentire (1973) cité par D. Holmes dans *Literary & Linguistic Computing* 13, 3, p. 112.
- 2.. L'historique de D. Holmes (art. cité) rappelle d'ailleurs que les premiers travaux de stylométrie ont porté sur la longueur des mots (p. 112). Parmi les articles récents, on se reportera par exemple à ceux de H. Baayen, H. Van Halteren & F. Tweedie (1996), de H. Somers & F. Tweedie (2003) et de E. Stamatos, N. Fakotakis & G. Kokkinakis (2001). Pour le latin, on consultera également les travaux de M. Dubrocard, notamment (2002).
- 3.. L'article d'Étienne Brunet, ici-même, rappelle ce constat récurrent qu'il convient de toujours garder à l'esprit avant d'émettre quelque hypothèse que ce soit sur la parenté supposée de deux ou plusieurs textes.
- 4.. Cf. par ex. les travaux de B. Pincemin.
- 5.. Rappelons que le critère de la longueur des phrases qui pourrait répondre aux mêmes exigences et qui est souvent présenté comme le plus efficace ou le plus rentable (meilleur pourcentage de capacité prédictive sur la variance) est malheureusement inutilisable pour les textes latins puisque ceux-ci ont été écrits à l'origine sans ponctuation (*scriptio continua*) et que seuls les éditeurs modernes sont responsables de la ponctuation qui les segmente actuellement.
- 6.. On lira cependant avec profit l'article de F. Rastier (2001) qui pose les premières bases d'une réflexion en la matière.
- 7.. C'est aussi la condition posée par E. Stamatos, N. Fakotakis & G. Kokkinakis (2001).
- 8.. Base élaborée et fournie par le Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) de l'Université de Liège. Pour une description plus précise, cf. S. Mellet (1996).
- 9.. C'est-à-dire substantifs, adjectifs, pronoms autres que relatifs, verbes, subordinants regroupant les relatifs et les conjonctions de subordination, adverbes, autres.
- 10.. Le latin connaît deux nombres comme le français (singulier et pluriel) et sept cas : nominatif, vocatif, accusatif, génitif, datif, ablatif et locatif, qui permettent d'exprimer les différentes fonctions du nom dans la proposition.
- 11.. Aux classiques indicatif, impératif, subjonctif, participe, gérondif et infinitif, il faut ajouter en latin un adjectif verbal et un supin. Les temps sont approximativement les mêmes qu'en français.

12.. Signalons qu'il existe d'autres calculs de distances qui tiennent compte des fréquences, par exemple celle de Hamming (cf. X. Luong 1994 : 29). La méthode d'É. Brunet, en revanche, ne prend pas en compte l'inégale répartition des formes communes et, pour cette raison, semble devoir être réservée aux calculs de distance lexicale.

13.. Il s'agit des œuvres de César (*Guerre des Gaules* et *Guerre Civile*), de trois de ses lieutenants (*Guerre d'Afrique*, *Guerre d'Espagne* et *Guerre d'Alexandrie*), d'une œuvre de Salluste (*Guerre de Jugurtha*), de quelques livres des *Annales* de Tacite auxquels s'ajoute son *de Vita Agricola*, et enfin de ce qu'il nous reste de l'*Histoire d'Alexandre* de Quinte-Curce. Pour un descriptif plus précis de ce corpus, voir l'article de D. Longrée et X. Luong ici-même.

14.. On ne peut donc plus parler, en toute rigueur, de « distance textuelle » ; il sera plus juste d'appeler ce calcul « mesure de distance grammaticale entre des textes ».

15.. Il y a bien sûr, en latin comme en français, des modes qui n'ont pas de temps (comme le gérondif par exemple) et des modes ou des verbes impersonnels ; d'où les variations numériques de ces sous-ensembles.

16.. Plus généralement, comme le remarque J. Burrows (2003 :7), toutes les méthodes se focalisent sur les phénomènes très fréquents, sur les traits linguistiques les mieux attestés plutôt que sur ceux qui sont rares.

17.. La réduction à ce sous-corpus de 11 textes a le mérite de neutraliser entièrement aussi bien la variable chronologique que la variable générique.

18.. Certains, comme H. Meusel, pensent même que le texte de la *Guerre Civile* n'est pas intégralement dû à la main de César. D'autres, face à la même disparité d'écriture, concluent simplement au caractère inachevé de l'œuvre et à l'extrême rapidité de sa rédaction.

19.. Rappelons le nom des œuvres dont la distance est ici calculée : *Guerre des Gaules* (GALL) et *Guerre Civile* (CIVI) de César, *Guerre d'Afrique* (AFRI), *Guerre d'Alexandrie* (ALEX), *Guerre d'Espagne* (HISP) (anonymes), *Guerre de Jugurtha* (JUGU) et *Catilina* (CATI) de Salluste, livres 3 à 5 et livres 7 à 9 des *Histoires* de Quinte-Curce (QC-5 et QC-9), *Annales* (ANNA) et *Histoires* (HIST) de Tacite.

20.. Il faut de toutes façons être conscient que les analyses en composantes principales, souvent utilisées dans ce but, ne fournissent pas non plus par elles-mêmes un test d'attribution d'auteur ; elles permettent seulement une évaluation de la ressemblance entre des textes et de la structuration d'un corpus. « Properly stated, the original question here is not 'Who is the author of X?', but 'Do the entries in this scatter-plot fall into any intelligible pattern?' » (J. Burrows, 2003 : 8).

21.. Voir aussi la conclusion de H. Somers & F. Tweedie (2003).

RÉSUMÉS

Le calcul de distance entre les textes a le plus souvent été effectué à partir du dénombrement des données lexicales ; nous nous proposons ici d'abord de tester la possibilité d'appliquer l'un des calculs disponibles de distance à des paramètres grammaticaux, puis de proposer notre propre

méthode à partir, non pas d'un tableau de contingences, mais d'un tableau de classement ordinal. Les textes soumis au calcul sont des textes latins empruntés à un corpus lemmatisé et étiqueté. Les différents résultats sont comparés et leur pertinence est évaluée au regard d'un savoir philologique préalable.

Measure of the grammatical distance between the texts

The calculation of intertextual distance is generally performed by studying lexical parameters. We will in the first place examine whether it is possible or not to apply one of the available methods to grammatical parameters, then we explain our own method, based on an ordinal classification table rather than a multiple contingency table. To present this methodology, we use Latin texts extracted from a lemmatized and tagged corpus. The different results will be compared and evaluated.

INDEX

Mots-clés : distance intertextuelle, lemmatisation, catégories grammaticales, analyse arborée, latin, mesure ordinale

AUTEURS

SYLVIE MELLET

« Bases, Corpus et Langage », UMR 6039, UNSA