Les atouts multiples de la lemmatisation : l'exemple du latin

Sylvie Mellet[1], Gérald Purnelle[2]

[1]C.N.R.S. – UMR 6039 « Bases, Corpus et Langage » – BP 3209 – 06204 Nice Cedex 3 – France

[2]Centre Informatique de Philosophie et Lettres – Université de Liège – B-4000 Liège – Belgique

Abstract

Based on the long experience of LASLA (Liège) and its Latin text data bank, this paper aims to exemplify the usefulness of lemmatization and morphological analysis for text corpus studies (occurrences queries, statistical lists and tests). Lemmatization brings linguistic accuracy of data and results, morphological analysis opens up richer directions. Limits of this method are the lack of syntactical analysis (parsing).

Résumé

En se fondant sur la longue et riche expérience du LASLA (Liège) et sur la banque de textes latins qu'il a lemmatisée, il est ici question de montrer l'utilité que présentent la lemmatisation, mais aussi l'analyse morphologique des formes textuelles, pour les études de corpus textuels et linguistiques, notamment la recherche d'occurrences, les relevés statistiques et l'application de tests statistiques. Ces atouts sont, dans le cas de la lemmatisation, une précision linguistique des données manipulées et des résultats produits bien plus grande que dans l'exploitation des seules formes textuelles ; pour l'analyse morphologique, richesse de la valeur ajoutée et ouverture de perspectives. Plusieurs exemples illustrent non seulement les avantages, mais aussi les limites d'un codage purement lexical et morphologique, mais non syntaxique.

Mots-clés : lemmatisation, corpus étiqueté, catégories grammaticales, textes latins.

1. La lemmatisation

La lemmatisation du latin n'est pas une pratique récente, puisque, par exemple, les méthodes ici décrites ont été développées au LASLA (Laboratoire d'Analyse des Langues Anciennes, Université de Liège), dès le lendemain de sa fondation en novembre 1961. Elles ont permis la constitution progressive d'une banque de données textuelles lemmatisées et analysées qui atteint actuellement 1,7 millions de mots pour une vingtaine d'auteurs et 160 textes différents.

Rappelons que le lemme est l'étiquette associée à toute forme textuelle, identifiant le lexème auquel elle appartient et correspondant à la forme qui le représente dans un dictionnaire de référence.

La méthode du LASLA est semi-automatique[1] : un programme informatique, se fondant sur des lexiques de radicaux et de désinences (ou plutôt de bases et de finales), produit pour chaque forme d'un texte toutes les propositions de lemmes possibles, ainsi que toutes les analyses

[1] Pour une description de la méthode et de ses principes, cf. Évrard, 1962 et Denooz, 1978.

morphologiques possibles de la forme, pour chaque lemme auquel elle peut correspondre. Quelques exemples :

```
QUAE   QVI   1  46A   2
       QVI   1  46J   2
       QVI   1  46J   6
       QVI   1  46L   6
       QVIS  1  47A   2
       QVIS  1  47J   2
       QVIS  1  47J   6
       QVIS  1  47L   6
       QVIS  2  48A   2
       QVIS  2  48J   2
       QVIS  2  48J   6
       QVIS  2  48L   6

LEGI   LEGO  2  53A14
       LEGO  2  5C071
       LEX      13E00
```

```
SATIS  SATA     12N
       SATA     12O
       SATIS 1  26Z00
       SATIS 2  60000
       SERO  3  5CN44 1
       SERO  3  5CO44 1

VENIS  VENA     11N00
       VENA     11O00
       VENEO    56B11
       VENIO    54B11
       VENVM    12N00
       VENVM    12O00
```

Intervient à ce stade une opération purement manuelle : texte en main, un latiniste choisit la proposition de lemme et d'analyse morphologique contextuellement correcte pour chaque forme. Cette étape du travail est coûteuse en temps humain : les structures morphologiques, syntaxiques et stylistiques de la langue latine nous ont en effet jusqu'à présent écartés de l'énorme investissement en temps, en analyse linguistique et en développement algorithmique que représenterait le développement d'une désambiguïsation automatique un tant soit peu large. Mais le coût entraîné par la sélection manuelle est productif, puisque l'on dispose pour tout texte, après sélection, d'un fichier plat dans l'ordre du texte à raison d'un enregistrement structuré par forme textuelle, où figurent de façon univoque toutes les informations qui y sont attachées : son lemme, sa référence textuelle, son numéro d'ordre dans le texte, son analyse morphologique complète (partie du discours, conjugaison, déclinaison, classe d'adjectifs, type de pronom ou d'adverbe, cas, nombre, voix, mode, temps, personne, degré de comparaison, genre des adjectifs et pronoms), à laquelle s'ajoute un codage manuel de la subordination des verbes (cf. *infra*). Exemple (Cicéron, *Pro Caecina*, 82 : *quoniam satis recusavi, venio iam quo vocas*).

```
&QVONIAM       QUONIAM          DA082000100100182014          001 0804834841
&SATIS        2SATIS            DA082000100200260000          001 0804941567
&RECVSO        RECUSAVI         DA082000100300351A14-  PX     001 0805035852
&VENIO         VENIO            DA082000100400454A11&         001 0805169636
&IAM           IAM              DA082000100500560000          001 0805262509
&QVO          2QUO              DA082000100600666011          001 0805333936
&VOCO          VOCAS            DA082000100700751B11-  ND   2001 0805475479
```

2. Les atouts en termes de recherches d'occurrences

On s'avise, devant les listes de propositions, du haut degré d'amphibologie que peut atteindre la langue latine, tant sur le plan lexical (une forme peut appartenir à plusieurs lemmes) que sur le plan morphologique (une forme peut souffrir plusieurs analyses, y compris à l'intérieur d'un même lemme). Il n'est pas besoin de décrire davantage l'un des principaux atouts d'une méthode semi-automatique, qui ne se limite pas à la lemmatisation, mais induit de manière automatique toutes les analyses morphologiques d'une forme, et donc la bonne. Cette méthode, qui, historiquement, remonte aux débuts de l'application de l'informatique aux études linguistiques et littéraires, n'a pas dû évoluer ou se modifier dans son détail, et l'on notera que

ce qui existe à l'heure actuelle pour les langues modernes n'en diffère guère sur le plan méthodologique (p. ex. Intex).

La notion de lemme est évidemment capitale. Elle permet d'atteindre et de regrouper toutes les occurrences d'un lexème, notion plus stable et linguistiquement plus cohérente que celle de forme textuelle. Dans un texte non lemmatisé, essayer de reconstituer une telle liste par la recherche de toutes les formes possibles d'un mot peut s'avérer fastidieux, voire dangereux, alors que la lemmatisation permet de rapporter automatiquement à leur lemme des variations morphologiques ou graphiques parfois très grandes ; quelques exemples latins : les lemmes *sum* « être » ou *fero* « porter » présentent plusieurs thèmes (*s-*, *er-*, *fu-* ; *fer-*, *tul-*, *lat-*) ; dans les formes des verbes à préverbes, l'assimilation consonantique est ou n'est pas attestée (*adcurro*, *accurro*). De même, l'homonymie des formes (l'appartenance de nombre d'entre elles à plusieurs lemmes) est résolue par la lemmatisation (cf. *supra* pour des formes comme *legi*, *satis* ou *venis*), qui exclut la production de bruit.

La banque de données constituée par les fichiers ainsi produits se prête à de multiples exploitations : production d'*indices verborum* et de concordances imprimés, concordances sélectives, requêtes portant sur un lemme ou un critère grammatical, relevés d'occurrences de phénomènes linguistiques, listes de fréquences décroissantes, relevés statistiques, etc. Pour cette exploitation, des instruments spécifiques ont été développés récemment : un logiciel de consultation et d'extraction (Estela, Nice) ; un site Web avec interface de requête et de consultation en ligne (LASLA) ; enfin les données du LASLA viennent d'être adaptées au programme Hyperbase d'Étienne Brunet.

Nous énumérons d'abord les diverses possibilités de recherches simples, avant de passer à des exemples d'utilisation et de consultation plus complexes.

Chaque étiquette attachée à une forme dans un fichier lemmatisé et analysé peut se prêter à une requête : non seulement le lemme (y compris la recherche de tous les lemmes présentant le même début, ou la même fin), mais aussi tous les codes morphologiques détaillés plus haut, en éventuelle combinaison avec le lemme. Concrètement, les objets de requête possibles sont les suivants : un lemme ; une forme ; tout lemme (ou toute forme) commençant ou finissant de la même manière ; tout critère morphologique, simple (tous les verbes au subjonctif) ou complexe (tous les verbes au subjonctif présent, tous les substantifs à l'accusatif, etc.) ; un type de subordination (tous les verbes introduits par le relatif *qui*) ; la combinaison d'un lemme (ou d'une forme) avec un critère morphologique (le verbe *fero* au passif, le substantif *pater* au pluriel) ou syntaxique (le verbe *facio* introduit par la conjonction *cum*) ; la combinaison des critères morphologique et syntaxique (tous les verbes au subjonctif introduit par la conjonction *cum*).

Un exemple de cumulation de critères, avec une contrainte sur la classe de mots, montre clairement combien lemmatisation et codage morphologique permettent d'éviter le bruit dans certains résultats. Les lemmes se terminant en *–ter* sont fréquents ; ainsi, dans l'*Énéide* de Virgile, on compte 45 lemmes différents, pour 6 catégories différentes et 993 occurrences, dont la répartition se fait comme suit :

Catégories	Occurrences	Lemmes	Catégories	Occurrences	Lemmes
Substantifs	460	15	Adverbes	81	13
Adjectifs	127	8	Adverbes numéraux	43	2
Pronoms	156	3	Prépositions	126	4

Substantifs : *accipiter* (1 occ.), *adulter* (1), *Auster*(13), *crater* (7), *culter* (1), *frater* (40), *iter* (35), *Iuppiter* (86), *magister* (12), *mater* (84), *minister* (3), *oleaster* (1), *pater* (172), *raster* (2), *venter* (2) ; adjectifs : *ater* (71), *dexter* (12), *equester* (1), *exter* (32), *pedester* (2), *sequester* (1), *sinister* (6), *teter* (2) ; pronoms : *alter* (29), *noster* (94), *vester* (33) ; adverbes : *acriter* (1), *aliter* (14), *breviter* (9), *crudeliter* (1), *graviter* (8), *licenter* (1), *molliter* (2), *ociter* (10), *pariter* (30), *praeter* (1), *segniter* (1), *subter* (1), *suppliciter* (2) ; adverbes numéraux : *quater* (5), *ter* (38) ; prépositions : *inter* (116) *praeter* (1) *propter* (4) *subter* (5).

Sans le codage morphologique, celui qui s'intéresse aux mots en *–ter* d'une catégorie particulière (par exemple les adjectifs) se heurterait à un bruit abondant, quelle que soit la catégorie prise en compte. Sans la lemmatisation, il serait contraint de rechercher toutes les finales flexionnelles possibles, en générant ainsi un bruit plus abondant encore.

3. Les requêtes sur critères croisés

Nous nous étendrons davantage sur le croisement des critères, notamment des critères grammaticaux. Ceci nous permettra de montrer non seulement les atouts, mais aussi les limites du codage du LASLA.

3.1. Listes fréquentielles

À côté d'un index fréquentiel totalement fiable (aucune occurrence oubliée, aucun homonyme confondu), on peut encore créer des sous-index selon la classe de mots : on obtiendra ainsi sans difficulté les 30 substantifs les plus fréquents d'une œuvre ou ses 50 premiers verbes — et eux seuls. Les calculs de spécificités et les analyses thématiques s'en trouvent affinées, comme le suggère l'exemple suivant : nous avons établi l'index fréquentiel décroissant de deux œuvres poétiques, les *Géorgiques* de Virgile et les *Élégies* de Tibulle ; ces index, au premier abord, ne révèlent rien d'autre que la thématique spécifique attendue pour chacune des œuvres : hymne à la terre et à la vie rurale pour la première, hymne à l'amour pour l'autre. Cependant, la projection sur ces listes fréquentielles d'un critère de tri par classes de mots (ou parties du discours) révèle que la différence thématique est essentiellement portée par les substantifs, alors que verbes et adjectifs sont beaucoup moins spécifiques : en effet sur les 24 lemmes adjectivaux les plus fréquents dans chaque œuvre, 10 sont communs aux deux listes (soulignés ci-dessous), soit près de 42 %. Sur les 24 lemmes substantivaux les plus fréquents, 5 seulement sont communs. Hasard, pauvreté de la langue latine en matière de description, plus grande polysémie des adjectifs ? Ce n'est pas le lieu ici de proposer une interprétation. On souhaite simplement montrer que le codage grammatical ouvre dans ce domaine des pistes de réflexion nouvelles et qu'il ne répond pas exclusivement aux besoins du grammairien : les lexicologues et stylisticiens peuvent y trouver leur compte également.

Tibulle, *Élégies*			**Virgile, *Géorgiques***		
Il y a 12 635 mots et 2 626 lemmes différents			Il y a 14 834 mots et 3 470 lemmes différents		
Il y a 538 éléments <adjectif>			Il y a 684 éléments <adjectif>		
Parmi les 538 éléments <adjectif> il y en a 24 parmi les 200 lemmes les plus fréquents			Parmi les 684 éléments <adjectif> il y en a 24 parmi les 200 lemmes les plus fréquents		
41 MAGNVS	19 SACER	14 DIVES	61 MAGNVS	23 DVLCIS	15 EXTER
32 TENER	17 CANDIDVS	14 DVLCIS	48 ALTVS	22 TENVIS	15 GRAVIS
31 MVLTVS	17 SAEVVS	13 BONVS	37 MVLTVS	21 MOLLIS	15 LEVIS
25 TRISTIS	16 FELIX	13 FERVS	33 LAETVS	21 SVPERVS	14 LONGVS

22 DVRVS	16 LEVIS	13 MOLLIS	31 INGENS	20 DENSVS	14 TRISTIS
22 LONGVS	16 PARVVS	13 NIVEVS	28 PINGVIS	19 NIGER	13 ACER
22 MISER	14 CARVS	12 CASTVS	26 DVRVS	19 NOVVS	13 VARIVS
20 SANCTVS	14 CELER	12 NIGER	25 MEDIVS	16 FRIGIDVS	12 BONVS
Parmi les 809 éléments <substantif> il y en a 55 parmi les 200 lemmes les plus fréquents			Parmi les 1 194 éléments <substantif> il y en a 78 parmi les 200 lemmes les plus fréquents		
68 DEVS	27 AQVA	22 VIA	61 TERRA	30 FLVMEN	24 PECVS
52 PVELLA	27 CVRA	21 CAPVT	45 CAELVM	28 OS	24 TECTVM
39 AMOR	27 IVVENIS	20 VIR	39 SILVA	27 CORPVS	23 NOX
36 PES	25 PVER	20 VNDA	34 LABOR	27 VNDA	23 VMBRA
33 MANVS	24 CARMEN	19 ANNVS	33 HERBA	26 ARVVM	21 VIS
32 TERRA	24 DIES	19 VOTVM	33 SOL	26 VENTVS	20 AGER
30 VERBVM	24 DOMVS	18 AGER	31 CAMPVS	25 ARBOR	20 ANIMVS
29 COMA	24 NOX	18 ARS	31 CVRA	25 IGNIS	20 DEVS

3.2. Recherche de cooccurrences

La recherche de cooccurrences est une autre façon de croiser paramètres lexicaux et paramètres grammaticaux : on peut en effet envisager, à côté de la recherche de deux formes ou de lemmes cooccurrents, la recherche d'un lemme et d'une catégorie grammaticale, voire de deux catégories grammaticales. Voici un exemple pour chaque cas de figure :

Recherche, dans les *Catilinaires* de Cicéron, des cooccurrences d'un impératif (présent ou futur, actif ou passif) et d'un vocatif, c'est-à-dire de la forme nominale qui exprime en latin l'apostrophe :

Recherche de cooccurrences :

1er élément : verbe
Conjugaisons > Toute conjugaison
Voix > Toutes
Modes > Impératif
Personnes > Toutes
2nd élément : substantif
Déclinaisons > Toutes
Cas > Vocatif
Nombre > Singulier & Pluriel

Résultats dans les deux premiers livres des Catilinaires *de Cicéron :*

I, 10 : quae cum ita sint, Catilina, perge quo coepisti, egredere aliquando ex urbe ! « Puisque les choses sont ainsi, Catilina, poursuis ce que tu as commencé, sors une bonne fois de la ville. »
I, 10 : quae cum ita sint, Catilina, perge quo coepisti, egredere aliquando ex urbe ! « Puisque les choses sont ainsi, Catilina, poursuis ce que tu as commencé, sors une bonne fois de la ville. »
I, 20 : egredere ex urbe, Catilina, libera rem publicam metu ! « Sors de la ville, Catilina, libère la république de la peur. »
I, 20 : egredere ex urbe, Catilina, libera rem publicam metu ! « Sors de la ville, Catilina, libère la république de la peur. »
II, 24 : instruite nunc Quirites contra has tam praeclaras Catilinae copias uestra praesidia uestrosque exercitus (…) « Disposez donc, citoyens, contre ces brillantes troupes de Catilina, vos garnisons et vos armées. »

Recherche, dans un livre de la *Guerre des Gaules* de César, de *dum*, conjonction de subordination suivie de l'indicatif présent, en cooccurrence — dans une limite de quatre mots — avec un pronom démonstratif ou anaphorique :

Recherche de cooccurrences :

1er élément : subordonnant *DVM*
suivi de l'indicatif présent
2nd élément : pronom
Nature > Démonstratif

Résultats dans le livre 7 de la Guerre des Gaules *de César :*

37,1 : dum haec ad Gergouiam geruntur, […].
42,1 : dum haec ad Gergouiam geruntur, [...].
57,1 : dum haec apud Caesarem geruntur, […].
66,1 : interea dum haec geruntur, [...].
75,1 : dum haec ad Alesiam geruntur, [...].
82,3 : at interiores, dum ea quae ad eruptionem praeparauerant proferunt, priores fossas explent, diutius in his rebus administrandis morati prius suos discessisse cognouerunt quam munitionibus adpropinquarent.

Ce type de requête soulève cependant quelques difficultés techniques auxquelles on ne pense pas toujours de prime abord : une option banale dans un programme de recherche de cooccurrences consiste en effet à proposer à l'utilisateur le choix de l'ordre des deux éléments en cooccurrence et une limitation possible de leur distance en nombre de mots. Or ce dernier point est ici plus délicat à gérer qu'il n'y paraît : il doit pouvoir se combiner avec l'instruction de recherche au sein d'une même phrase et d'un même paragraphe pour garder toute sa valeur morpho-syntaxique ; en effet, autant la répétition d'un même vocable à quelques mots d'intervalle peut faire sens (en termes stylistiques, sémantiques ou de cohésion textuelle par exemple) même lorsque les deux occurrences sont séparées par une ponctuation forte, autant une telle configuration paraît peu pertinente lorsqu'on travaille sur les cooccurrences de deux catégories grammaticales (conjonction et mode verbal, préposition et cas, interjection et impératif, etc.). Or une telle contrainte oblige non seulement à complexifier le programme de requête pour compléter la définition de la distance en nombre de mots par la prise en compte des ponctuations fortes et des espacements de paragraphes, mais elle nécessite aussi de pouvoir différencier les points d'abréviation des points de ponctuation et, en poésie, les retours à la ligne marquant la fin de vers de ceux qui marquent la fin de paragraphe ; car la phrase, unité de localisation de la cooccurrence, peut bien évidemment s'étendre sur plusieurs vers. Ce problème doit donc être pris en compte dès le codage des données. Dans les fichiers du LASLA, les fins de phrases sont codées dès la production de la référence.

Enfin on notera que la recherche de cooccurrences est un palliatif — insatisfaisant, il faut le reconnaître — à l'absence de codage des liens de dépendance syntaxique. En effet dans notre base les liens syntaxiques n'ont été codés que dans une seule configuration, celle d'un verbe subordonné dépendant d'une conjonction de subordination (celle-ci fût-elle le morphème zéro dans le cas d'une dépendance paratactique). Il est à noter que c'est bien le verbe qui est codé, comme régi et dépendant de tel ou tel subordonnant, et non pas la conjonction. Ce procédé, transposable naturellement à d'autres langues que le latin, présente des avantages considérables : il permet de résoudre très facilement le cas où plusieurs subordonnées sont coordonnées sans que le subordonnant lui-même soit répété ; ainsi relève-t-on, sans difficulté aucune, deux propositions temporelles dans la phrase « Quand il ouvrit la porte et partit, je poussai un soupir de soulagement ». Par ailleurs, ce procédé de codage permet un classement rapide des subordonnées en fonction du temps et du mode de leur verbe : c'est en effet sur la

même et unique étiquette associée au verbe subordonné que le programme d'exploration doit rechercher la nature de la subordonnée, le temps et le mode verbal. On devine le gain de temps par rapport à une recherche qui devrait d'abord lister les occurrences de la conjonction choisie, puis vérifier et classer la forme verbale associée.

Pour tous les autres cas de dépendance ou de lien anaphorique en revanche, on se trouve assez démuni (accord adjectif / substantif, verbe / sujet, relatif / antécédent, etc.). Rien n'indique le statut de terme régi et de terme régissant et tout se passe comme si « le discours, qui se déroule dans le temps, n'[avait] qu'une dimension et [était] irréversible, si bien que les connexions possibles pourraient paraître limitées à la séquence immédiate, soit progressive, soit régressive » (Évrard, 1989 : 116). Bref, pour le dire vite avec des anglicismes jargonnants, le LASLA a eu le courage et les moyens de « tagger » ses textes, mais pas de les « parser » ! Voici un exemple dans lequel on voit clairement que la recherche de cooccurrences ne compense qu'imparfaitement cette faiblesse du codage initial : le participe passé passif latin connaît deux types principaux de construction de son complément d'agent lorsque celui-ci est exprimé : le datif seul ou l'ablatif précédé de la préposition *ab* (ou *a*). Une recherche automatique de ces deux formes de complémentation peut bien sûr se faire par recherche de cooccurrences ; mais les résultats sont bruités, comme le montrent les relevés suivants :

Recherche de cooccurrences :

1er élément : lemme *AB*
2nd élément : verbe :
Conjugaisons > toute conjugaison
Voix > passive
Modes > participe
Temps > parfait
Cas > Tous
Ordre des éléments : indifférent ; distance : 5 mots

Résultats corrects dans Salluste, Guerre de Jugurtha *:*

5,4 : [...] Masinissa, rex Numidarum in amicitiam receptus a P. Scipione, […] « Massinissa, roi des Numides reçu par Scipion dans notre amitié […]. »
11,4 : Dein tamen ut aetati concederet fatigatus a fratre […] « Pourtant, harcelé par son frère [qui le pressait] de s'incliner devant l'âge […]. »
Etc.

Mais aussi :

14,5 : Ceteri reges […] bello uicti in amicitiam a uobis recepti sunt [...] « Les autres rois, une fois vaincus à la guerre, ont été reçus par vous dans votre amitié. »
18,9 : Nam, freto diuisi ab Hispania, mutare res inter se instituerant « Car, n'étant séparés de l'Espagne que par un détroit, ils avaient établi avec ce pays des échanges commerciaux. »

La conclusion évidente à tirer de ces remarques est que toute économie de moyens au moment de la constitution et de l'annotation d'une base de données textuelles se retrouve nécessairement dans les capacités de traitement ultérieur ; et que l'ajout *a posteriori* d'étiquettes ou de liens est fastidieux, voire pratiquement impossible.

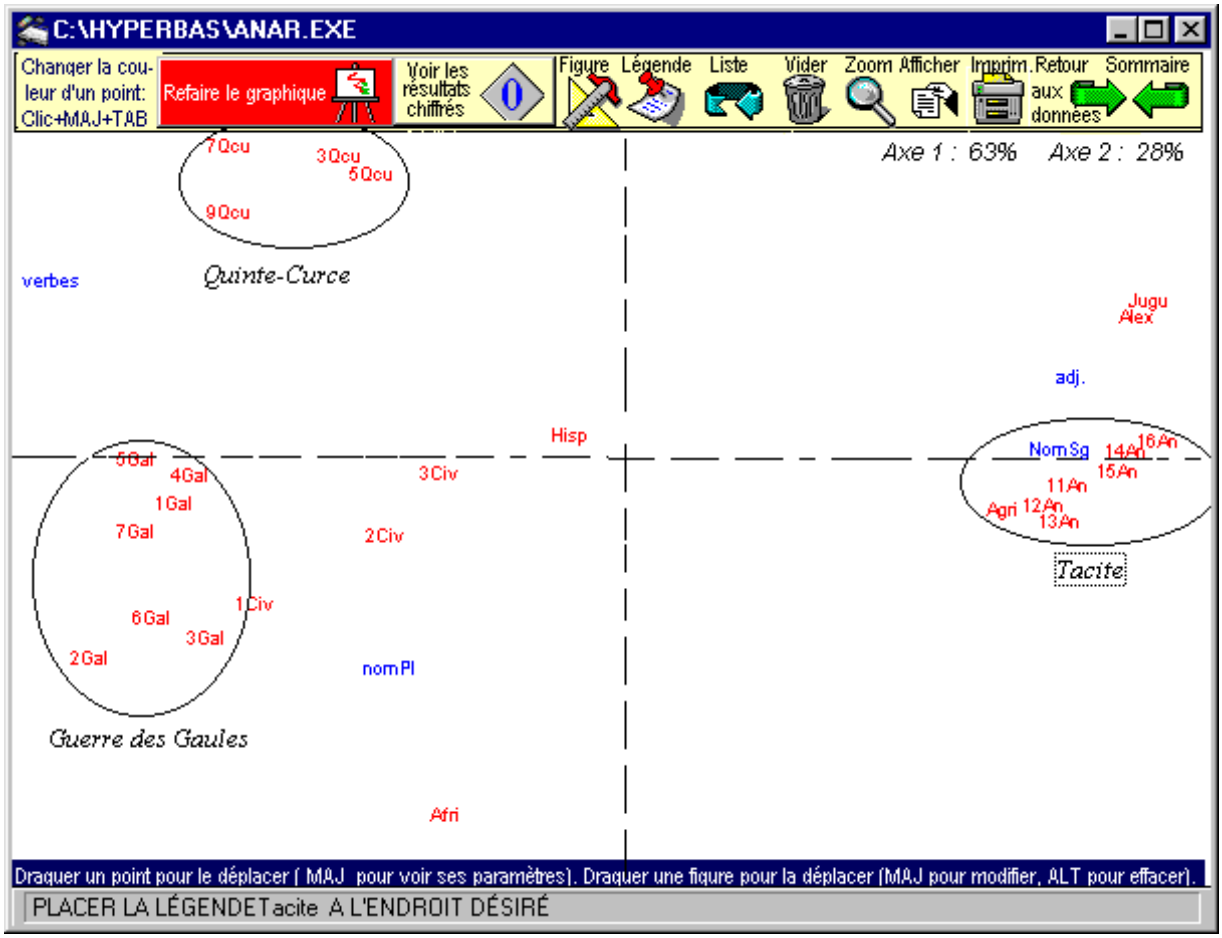
4. Les traitements statistiques

Une des objections fortes à la rentabilité de la lemmatisation a été formulée par Étienne Brunet qui montre, de manière indiscutable, que l'un des traitements statistiques de base, à savoir le calcul de la connexion lexicale entre différents textes (ou calcul de distance intertextuelle) donnait des résultats sensiblement identiques, qu'on le fasse à partir des formes ou à partir des

lemmes (Brunet, 2000 et ici-même). De fait, sur notre corpus latin, les regroupements par auteurs et par genres littéraires se font exactement de la même façon dans les deux cas ; c'est certainement un point qui mérite attention et, peut-être, approfondissement théorique (que mesure exactement la connexion lexicale, qui est si peu sensible à la variation des formes du vocabulaire ?).

On se contentera ici de deux remarques pour amorcer la discussion :

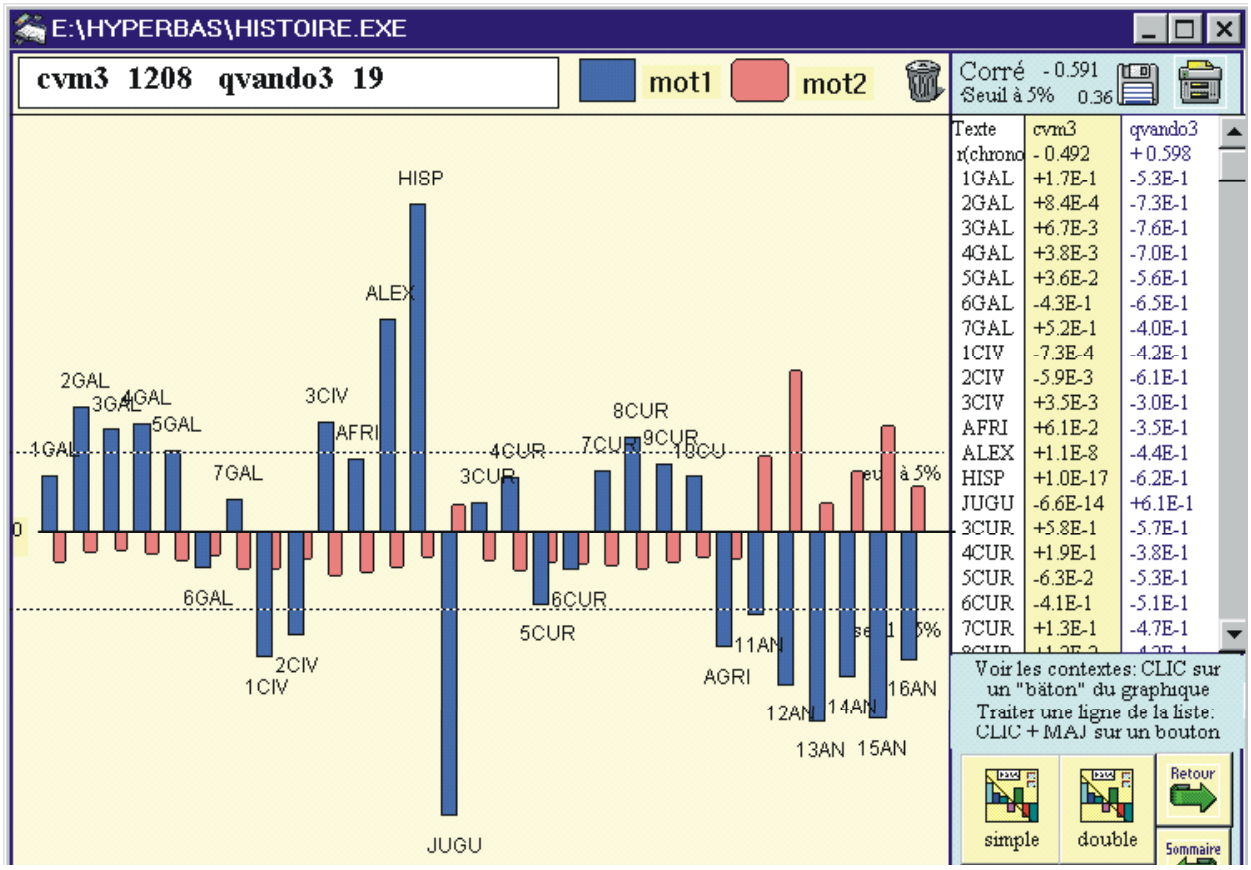
4.1. D'une part les calculs de distance peuvent gagner à être enrichis par de nouveaux paramètres ; dit autrement, la connexion lexicale n'est pas la seule forme de connexion possible : Sylvie Mellet travaille actuellement avec Xuan Luong à l'élaboration d'une méthode mathématique adaptée au calcul de la connexion grammaticale (Luong et Mellet : à paraître). En attendant les résultats de cette recherche, voici simplement une analyse factorielle portant sur la fréquence respective de quelques classes de mots chez les historiens latins :



Les regroupements sont ici remarquables : l'œuvre de Tacite forme bien un tout homogène (y compris la *Vie d'Agricola*), centrée autour de l'emploi de substantifs singuliers, ce qui correspond à un style plus abstrait que celui des autres historiens. L'œuvre de Quinte-Curce est tout aussi clairement regroupée, autour du verbe cette fois. Chez César, seule la *Guerre des Gaules*, accompagnée peut-être du premier livre de la *Guerre civile*, constitue un ensemble aussi homogène. Les autres livres de la *Guerre civile* manifestent une tendance à la dispersion qui confirme les observations faites par la critique philologique traditionnelle : il est probable en effet que, dans ces livres, de larges passages ne sont pas de la main de César. Quant aux trois œuvres attribuées à des lieutenants de César, elles se répartissent elles aussi de manière fort intéressante : deux d'entre elles (*Guerre d'Espagne* et *Guerre d'Afrique*) imitent sans

doute assez bien le style du maître, sans se confondre toutefois avec lui ; mais la troisième, la *Guerre d'Alexandrie*, s'en démarque nettement par un suremploi de l'adjectif qui la rapproche étonnamment de la *Guerre de Jugurtha* racontée par Salluste.

4.2. Autre remarque, triviale : la statistique linguistique (et, par la même occasion, les fonctionnalités du logiciel HYPERBASE) ne se réduisent pas à l'artillerie lourde du calcul de distance intertextuelle. D'autres tests, tout aussi intéressants pour le linguiste, voient leur intérêt s'accroître considérablement avec la lemmatisation et le codage grammatical. En voici un seul exemple : il s'agit d'un test fait dans le cadre d'une recherche sur l'évolution diachronique de la subordination en latin ; on a voulu comparer la distribution, dans un corpus d'historiens, de deux conjonctions partiellement concurrentes : *quando* qui signifie « quand » et *cum* qui signifie « comme, alors que ». Or il faut savoir que *quando* peut être aussi en latin un indéfini et un interrogatif (« un jour », « quand ? ») tandis que *cum* peut être aussi préposition (« avec ») : sans lemmatisation, impossible de distinguer ces homonymes et impossible de comparer seulement les deux conjonctions de subordination ; c'eût été dommage car les résultats révèlent une corrélation inverse fort intéressante :



On espère, à travers ces applications, avoir montré la pertinence de la lemmatisation et du codage grammatical de notre corpus latin, suggéré des pistes de recherche parallèles dans le domaine des langues vivantes et contribué au débat sur la rentabilité de cette lemmatisation.

Références

Brunet É. (1999). *Logiciel HYPERBASE*, hypertexte statistique pour grands corpus, nouvelle version Mac et Windows et manuel de référence. Nice : INaLF / Paris : Champion.

Brunet É. (2000). Qui lemmatise attise. In José, L. et Theissen, A. éditeurs, *Scolia*, n 13 (Actes des 11es rencontres linguistiques en pays rhénan), Strasbourg, pp. 7-32.

Denooz J. (1978). L'ordinateur et le latin, techniques et méthodes. *RELO*, vol. 4 : 1-36.

Denooz J. (1998). Aperçu des travaux du LASLA. *RISSH*, vol.34 : 51-70.

Évrard É. (1962). Le laboratoire d'analyse statistique des langues anciennes de l'Université de Liège. *Mouvement scientifique en Belgique*, vol. : 163-169.

Évrard É. (1989). Une informatisation de la syntaxe de dépendance en latin. In Lavency, M. et Longrée, D. éditeurs, *Actes du Ve Colloque de Linguistique latine*, Louvain-la-Neuve, Peeters (CILL 15), pp. 115-126.

Habert B., Nazarenko A. et Salem A. (1997). *Les linguistiques de corpus*. Paris, A. Colin.

Habert B., Fabre C. et Issac F. (1998). *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques*. Paris, InterÉditions / Masson.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble, Cahiers du CERAT.

Mellet S. (1994). Logiciels d'exploitation de la banque de données de textes latins du L.A.S.L.A. *RISSH*, vol. 30 : pp. 91-108.

Mellet S. (1996). Les atouts de la lemmatisation. In Moracchini G. éditeur, *Bases de données linguistiques : conceptions, réalisations, exploitations*. Univ. de Corse / Univ. de Nice – Sophia Antipolis, pp. 309-316.

Purnelle G. (1996). Utilisation d'une banque de données de textes latins lemmatisés et analysés. Problèmes spécifiques aux données linguistiques. In Moracchini G. éditeur, *Bases de données linguistiques : conceptions, réalisations, exploitations*. Univ. de Corse / Univ. de Nice – Sophia Antipolis, pp. 295-307.