Université de Corse - Université de Nice Sophia Antipolis

Bases de données linguistiques : conceptions, réalisations, exploitations

Actes du Colloque International de Corte
(11-14 octobre 1995)

Textes réunis et édités par Georges Moracchini

Sylvie MELLET
UPRES-A «Bases, corpus et langage»
CNRS-INaLF

LES ATOUTS DE LA LEMMATISATION

C'est à partir d'une banque de données de textes latins que je voudrais évoquer les atouts de la lemmatisation ; le point de départ de ma réflexion se situe donc un peu en marge des principaux centres d'intérêt de ce colloque, mais je tenterai d'en montrer la portée générale et d'en faire un exemple transposable à d'autres domaines linguistiques. La banque de données en question est le fruit du travail du Laboratoire d'Analyse Statistique des Langues Anciennes de l'université de Liège (L.A.S.L.A.)[1] et je voudrais insister, au cours de cette communication, sur une spécificité importante de cette réalisation : dès sa conception, dans le début des années soixante, cette banque textuelle a été structurée par le concept de lemmatisation. On pourrait penser que, s'agissant d'une banque de données consacrée à une langue flexionnelle, le parti pris d'une telle lemmatisation devait s'imposer d'évidence comme une nécessité. Tel n'a pourtant pas été le cas. D'autres équipes de recherche en langues anciennes ont fait le choix d'enregistrer les textes latins à l'état brut, arguant que le coût de la lemmatisation - notamment en temps - était excessif par rapport à ses avantages. La banque du L.A.S.L.A. reste donc encore à l'heure actuelle, à ma connaissance, la seule à avoir été d'emblée entièrement lemmatisée, même s'il est vrai que certaines équipes (par exemple à l'université de Turin) travaillent à des programmes de lemmatisation auto-

matique du latin. C'est donc d'abord sur les principes, puis sur les atouts de cette lemmatisation que je vais centrer mon propos.

1. Les principes de la lemmatisation

Le principe fondamental consiste, bien sûr, à se donner les moyens de reconnaître un même lemme à travers toutes ses occurrences dans un texte, quelle que soit la variété des formes attestées à travers ces occurrences. Or ces variations sont très nombreuses en latin ; elles tiennent à deux particularités de la langue :

- au fait que le latin est une **langue flexionnelle** où non seulement le verbe est conjugué, mais où le nom et l'adjectif sont déclinés (*rosa, rosam, rosae, rosa, rosas, rosarum, rosis*) ; les adjectifs varient bien sûr en genre (il y en a trois en latin), en nombre, en cas (six) mais ont aussi le plus souvent un comparatif et un superlatif synthétique (*doctus, doctior, doctissimus :* " savant ", " plus savant ", " très savant "). La forme même du radical peut être sensiblement modifiée par la flexion casuelle : *homo/hominis* " l'homme ", *iter/itineris* " le chemin ", *genus/generis* " l'espèce "; enfin les verbes n'ont pas besoin d'être irréguliers pour connaître des variations de radical importantes dans leur conjugaison : *tango/tetigi/tactus* (présent, parfait, participe passé du verbe " toucher ").
- au fait que **l'orthographe latine n'est pas entièrement stabilisée** ; on observe donc au sein d'un même texte parfois ou entre textes différents de même époque des variantes orthographiques importantes, traduisant des hésitations sur l'oportunité de transcrire les assimilations consonantiques (*inlicio/illicio* ; *adtuli/attuli* ; *quidquid/quicquid*), les haplologies (*exspecto/expecto*) ou trahissant la faiblesse du statut phonologique de certains phonèmes (*h-* initial, semi-initial ou intervocalique : *harena/arena, exhibeo/exibeo, mihi/mi* ; *n-* devant sifflante ou fricative : *consul/cosul* ; etc.)

On a pu penser qu'il s'agissait là de traits spécifiques du latin imposant des contraintes qu'on ne retrouvait pas - ou moins - ailleurs, en particulier en français : je me souviens de discussions avec des francisants estimant, il y a une dizaine d'années, que, pour les textes français, la recherche sur formes graphiques était amplement suffisante si l'on voulait bien ne pas s'attaquer à la poignée de verbes irréguliers de notre lexique. Leur sentiment a en général bien évolué depuis. Et, de fait, les deux caractéristiques que nous venons d'attribuer au latin ne lui appartiennent pas en propre : parmi les langues proches génétiquement ou géographiquement le roumain, l'allemand, le basque déclinent

les substantifs et les adjectifs ; pour nous en tenir aux langues réputées non flexionnelles telles que le français ou l'italien, il faut bien reconnaître qu'elles ont une flexion verbale très riche et qu'elles conservent des traces de flexion pronominale. Quant aux variantes orthographiques, il est inutile d'insister sur leur réalité contemporaine en présence d'un auditoire largement composé de dialectologues. En français même, l'orthographe n'a été stabilisée qu'assez récemment et des recherches diachroniques dans la base Frantext par exemple obligent à prendre en considération le problème.

En bref, l'intérêt de la lemmatisation n'est pas réservé aux spécialistes des langues dites classiques, mais ce sont peut-être certains d'entre eux qui l'ont développée avec le plus d'efficacité et de constance.

Voici en deux mots comment. Chacune des formes du texte a été analysée et rattachée à une entrée de dictionnaire sous laquelle sont regroupées toutes les variantes graphiques et morphologiques d'un même lemme. L'analyse est faite de manière semi-automatique, grâce à un analyseur morphologique qui utilise d'une part un dictionnaire des bases lexicales du latin classique, d'autre part un dictionnaire des diverses " désinences " (au sens très large) nominales et verbales. Les ambiguités, liées le plus souvent aux homonymies désinentielles, sont levées par un philologue.

L'essentiel est de comprendre que le texte est conservé dans sa forme d'origine, avec ses traits spécifiques, y compris ses éventuelles graphies dialectales, mais qu'est créé parallèlement un index des lemmes classés par ordre alphabétique dans lequel est enregistré également le numéro d'ordre dans le texte de chacune des occurrences de chaque lemme : ainsi le va-et-vient du texte à l'index est aisé et rapide.

Extrait du fichier " Index des lemmes " du ***De Amicitia***

AMICVS 112 8092	AMICVS 112 8105	AMICVS 112 8115	AMICVS 112 8138
AMICVS 112 8343	AMICVS 1128371	AMICVS 22J 374	AMICVS 22J 678
AMICVS 22J 3436	AMICVS 22J 7926	AMITTO 5C 1354	AMITTO 5C 3509

On voit ici que le lemme est répété autant de fois qu'il apparaît dans le texte ; il est suivi d'un code d'analyse minimal qui permet de lever les ambiguités dues à l'homonymie (ici 112 signifie : premier lemme *amicus*, substantif de la deuxième déclinaison, et 22J caractérise le second lemme *amicus*, adjectif au superlatif : voir § 2.1.2) ; vient ensuite le numéro d'ordre de chaque occurrence dans le texte.

Reste ensuite à élaborer les programmes capables de travailler sur l'un et l'autre fichier selon le type de recherche demandée : si je cherche toutes les formes contractes *nil* du lemme *nihil*, j'aurai besoin de parcourir le texte ; si je cherche l'ensemble des occurrences du pronom signifiant " rien " en latin, je consulterai l'index des lemmes. Dans les deux cas je pourrai afficher les contextes d'apparitions et produire une concordance.

2. Les atouts de la lemmatisation

2.1. L'exploitation du fichier " index des lemmes "

2.1.1. Le principal atout de ce fichier - et le plus évident - est de permettre une recherche très rapide (la consultation d'un fichier classé par ordre alphabétique est bien sûr plus rapide que la lecture linéaire d'un texte) et entièrement fiable, c'est-à-dire à la fois **exhaustive et non bruitée**. On est ainsi assuré de créer des concordances complètes même si le terme pivot présente des variantes orthographiques qui lui attribuent des localisations alphabétiques assez éloignées les unes des autres (cf. l'exemple ci-dessus de *arena* et *harena*) ou des variations morphologiques importantes ; chacun sait en effet que la recherche sur chaîne de caractères initiale à laquelle on recourt souvent quand les textes ne sont pas lemmatisés n'est qu'un palliatif incertain, notamment dans le domaine verbal : nombreux sont les systèmes de conjugaison qui reposent sur des alternances vocaliques du thème (verbes forts en allemand par ex.) ; parmi les verbes les plus fréquents d'une langue, tels que les verbes " faire ", " avoir ", " être ", " aller ", etc., il arrive même que la structure consonantique du radical subisse des modifications en cours de flexion, notamment en raison des phénomènes diachroniques de supplétisme. La lemmatisation préalable du texte latin ayant permis de rapporter une fois pour toutes les diverses formes d'un même verbe à son entrée de dictionnaire, ces difficultés d'origine morphologique sont donc levées ; ainsi la recherche de toutes les occurrences du verbe *tango* dans le *de Ira* de Sénèque affichera en quelques secondes les formes *tangetur*, *tangebatur*, *tetigerunt* et *tactus erit* ; de même, une recherche sur le verbe *fero* " porter " dans les livres I et VII de la *Guerre des Gaules* de César relèvera des formes aussi variées que *ferendus*, *ferebat*, *ferre*, *ferri*, *ferretur*, *tulisse*, *laturos*.

Ce " lissage " de la lemmatisation n'est pas seulement utile aux recherches lexicologiques ; il est aussi particulièrement avantageux

pour les recherches morphologiques, par exemple pour les études de **morphologie dérivationnelle** portant sur les termes suffixés ; la recherche d'un même suffixe est en effet particulièrement délicate dans une langue flexionnelle puisque les suffixes sont affectés par les variations désinentielles ; plus grave encore, certains types de flexion verbale peuvent reposer sur une opposition entre thème suffixé et thème non suffixé : ainsi le suffixe verbal *-sco*, de valeur inchoative, n'est, en raison de son signifié propre, présent qu'aux temps de l'infectum (c'est-à-dire au présent, imparfait et futur) et disparaît dans les autres formes de la conjugaison (*cognosco* " j'apprends à connaître "/*cognoui* " je connais "/*cognitum* " connu " ; *nascor* " je nais "/*natus sum* " je suis né ")

2.1.2. Il faut encore souligner que la lemmatisation permet le prétraitement de quelques difficultés récurrentes. Les plus importantes sont bien sûr constituées par les cas d'**homographie occasionnelle et d'homonymie lexicale** : j'appelle homographie occasionnelle les cas, assez fréquents en latin, où deux lemmes parfaitement différenciés produisent deux formes identiques au cours de leur flexion ; on citera par exemple la forme *is* qui peut être une occurrence du nominatif masculin singulier du pronom anaphorique " celui-ci " ou la deuxième personne du singulier du présent de l'indicatif du verbe *eo* " aller " ; ou encore *eas* accusatif féminin pluriel du même pronom et deuxième personne du singulier du subjonctif présent du même verbe ; *legis*, génitif singulier de *lex* " la loi " et deuxième personne du singulier du présent de l'indicatif de *lego* " lire " ; *regis*, génitif singulier de *rex* " le roi " et deuxième personne du singulier du présent de l'indicatif de *rego* " diriger " ; etc.[2] L'analyse préalable à la lemmatisation, sous le contrôle d'un philologue, a permis de lever les ambiguités et d'affecter chaque forme du texte à son lemme de référence.

Lorsque ce sont les lemmes eux-mêmes qui sont homonymes, la lemmatisation a obligé à les différencier, par exemple par des indices choisis soit en fonction de leur catégorie grammaticale, soit - si celle-ci est la même pour les deux lemmes - en fonction de leur fréquence. Ainsi *amicus* est codé 1 lorsqu'il s'agit d'un substantif et 2 lorsqu'il s'agit d'un adjectif. Toute recherche dans le fichier des lemmes prendra en compte cette codification et proposera à l'utilisateur un choix entre les différents lemmes homonymes ; admettons par exemple que nous recherchions la forme *quando* " quand " dans le *de Breuitate uitae* de Sénèque : le programme de consultation du fichier " index des lemmes " constatera que la forme en question répond en latin à quatre analyses gram-

maticales différentes codées de 1 à 4 et répondra à l'utilisateur que, dans le texte choisi, il trouve :

	QVANDO	
Tous		15
Quando 2	Adv. inter.	12
Quando 3	Conj. sub.	1
Quando 4	Adv. indéf.	2

A l'utilisateur alors de sélectionner sur l'écran la ligne de son choix.

2.1.3. On notera enfin que chaque lemme étant répété dans le fichier autant de fois qu'il apparaît dans le texte, il est aisé d'en calculer automatiquement la fréquence et de transformer l'index alphabétique en index fréquentiel normal ou inverse.

2.2. L'exploitation de l'analyse morphologique

Comme je l'ai déjà dit, la lemmatisation en latin ne pouvait se concevoir sans une analyse morphologique et syntaxique complète et extrêmement rigoureuse de la forme en contexte. Il aurait donc été dommage de ne pas conserver les informations fournies par cette analyse grammaticale et de ne pas tenter de les intégrer à la base de données.

Ces analyses ont donc été transcrites sous la forme d'un code alpha-numérique qui a été associé, dans le fichier texte, au mot qu'il décrivait. Ainsi, à chaque forme du texte latin est affecté un code porteur de son analyse grammaticale et de tous les repérages nécessaires à sa localisation dans le texte (numéro d'ordre dans la phrase, dans le paragraphe, dans le texte). Par exemple, le début de la *Guerre des Gaules* se présente ainsi :

Gallia C1001000100100111A00 est C10010001002002E6C11
omnis C1001000100300348A00 3 diuisa C100100010040045CC14&2
in C1001000100500570300 partes C1001000100600613L00
tres C1001000100700731L00 3 quarum C1001000100800846M11 2
unam C1001000100900931C00 2 incolunt C1001000101001053L11-LN
Belgae C1001000101101111J00

Après les indications de localisation dans le texte qui occupent les quinze premières colonnes, la seizième colonne indique la partie du

discours (1 pour substantif, 2 pour adjectif .... 5 pour verbe), la dix-septième donne le type de déclinaison ou de conjugaison en le désignant par son numéro habituel, la dix-huitième indique le cas et le nombre des noms, la personne pour les verbes ; les deux dernières colonnes sont réservées, notamment, aux temps et modes verbaux.

Il suffit donc de balayer chacune des colonnes de ces codes pour rechercher une catégorie grammaticale : par exemple tous les adjectifs du texte ou seulement tous les adjectifs au comparatif ; ou bien encore tous les verbes ou seulement tous les imparfaits du subjonctif, etc. Là encore les problèmes d'homonymie ont été réglés une fois pour toutes : notamment l'homonymie fréquente en latin des désinences a été élucidée et l'analyse correcte est conservée dans le code associé à chaque forme.

Sont également réglés par avance la reconnaissance et le traitement des périphrases verbales. En latin comme en français, une simple reconnaissance de formes graphiques ne permet pas en effet de différencier les occurrences d'un même lemme verbal selon qu'il est auxiliaire ou qu'il a un sens plein : en latin le problème se pose surtout pour les formes du verbe " être " (*esse*), beaucoup plus rarement pour certaines formes du verbe " aller " (*ire*) ; de nouveau ce travail d'analyse morphologique a été préalablement effectué lors de la lemmatisation et enregistré dans les fichiers de la banque de données[3].

En outre la mention du caractère subordonné de certains verbes sous la forme d'un code de deux lettres (LN après le code de *incolunt*) permet de créer un fichier de ces codes classés par ordre alphabétique. On peut donc ainsi créer un programme de recherche automatique sur ce nouveau fichier qui permette de sélectionner un type de subordonnée ou mieux encore un type de subordonnée en association avec tel ou tel temps et mode verbal, par exemple toutes les relatives au subjonctif (cf. la communication de G. Purnelle).

Au total, la structuration de la base de données est donc la suivante : un fichier texte où chaque forme est suivie de son code de localisation et d'analyse ; un fichier index des lemmes classés par ordre alphabétique ; un fichier des codes de subordination. **Or cette structuration a été entièrement déterminée par le choix de la lemmatisation.**

La lemmatisation induit ainsi le développement d'outils qui permettent une exploitation des textes sous tous leurs aspects linguistiques : lexicologiques et morphologiques bien sûr, mais aussi syn-

taxiques, voire énonciatifs et pragmatiques à travers notamment la lecture des codes affectés aux formes et celle du fichier consacré aux subordonnées.

Notes

[1] Voir l'exposé de G. Purnelle, ici-même.

[2] En français, une recherche sur les noms des saisons sera nécessairement entravée par toutes les occurrences du participe passé du verbe " être ".

[3] Dans l'exemple ci-dessus, on note le code E au lieu de 5 pour la forme *est* et l'indication, à la fin du code accompagnant *diuisa*, du lien syntagmatique que cette forme de participe entretient avec le 2ème mot de la phrase, c'est-à-dire précisément l'auxiliaire.