ANCIEN ET MOYEN FRANÇAIS SUR LE WEB

Enjeux méthodologiques et analyse du discours

sous la direction de

PIERRE KUNSTMANN

FRANCE MARTINEAU

DANIELLE FORGET

Les Éditions David

Typographie et montage : Anne-Marie Berthiaume graphiste

Les Éditions David, 2003
1678, rue Sansonnet
Ottawa (Ontario) K1C 5Y7

Téléphone : (613) 830-3336
Télécopieur : (613) 830-2819
ed.david@sympatico.ca

Site internet : www3.sympatico.ca/ed.david/

Tous droits réservés. Imprimé au Canada.
Dépôt légal (Québec et Ottawa), 4e trimestre 2003

Le Conseil des Arts du Canada | The Canada Council for the Arts
ONTARIO ARTS COUNCIL
CONSEIL DES ARTS DE L'ONTARIO
Ottawa

La lemmatisation de textes d'ancien français : méthodes et recherches

MARTIN-DIETRICH GLESSGEN

Université de Zürich

L'ÉTIQUETAGE et la lemmatisation de textes anciens en écriture alphabétique possède déjà une certaine tradition dans le domaine de l'informatique appliquée à la critique textuelle. La formation de cette tradition s'est caractérisée par un écart de plus en plus grand entre celle-ci et les évolutions dans le domaine de la philologie traditionnelle sur papier. Une mauvaise communication, durant les deux ou trois dernières décennies, a empêché de faire connaître de façon précise les possibilités de l'informatique dans les cercles de philologues ; en même temps, les spécialistes de l'informatique n'ont été que partiellement informés de l'apparition des normes philologiques actuelles et des contraintes sévères qu'elles impliquent : la transmission complexe des textes anciens (divers manuscrits pour un seul texte, nécessité de corrections et de commentaires) crée, pour la lemmatisation, des problèmes sérieux, que ne connaissent pas les langues standard modernes.

L'importance croissante des études de corpus et des réalisations effectuées grâce à l'informatique, dans la dernière décennie du vingtième siècle, conduit à une synthèse

nécessaire de ces deux approches, qui sont enracinées dans des traditions scientifiques très différentes. Seule une telle synthèse peut mener à la formation d'outils d'étiquetage, de lemmatisation et même de parsing des textes anciens, qu'on puisse considérer comme des programmes standard et que les spécialistes de critique textuelle puissent utiliser sans appréhension. Dans cette perspective, la mise au point de ces instruments s'accompagnera d'une réflexion explicite sur les avantages qu'ils présentent pour l'informatique ou pour la philologie.

L'auteur de ces lignes a choisi d'établir un prototype de programme de lemmatisation conçu spécifiquement pour les besoins philologiques des textes médiévaux de langue romane. Ce projet, réalisé en collaboration étroite avec Matthias Kopp (Tübingen), pourrait, dans le meilleur des cas, conduire à des applications utiles pour les chercheurs travaillant sur les textes ; même si ce n'était pas le cas, il serait tout de même précieux d'en comparer les avantages avec ceux d'outils aux fonctions similaires. Je reconnais que cette méthode pourrait ne pas sembler très facile ou directe, mais je n'en connais pas d'autre pour obtenir la compétence technique et un terme de comparaison suffisamment fondé pour être soumis à une évaluation. La question qui m'intéresse est de savoir sur quels critères on peut affirmer qu'un outil de lemmatisation particulier, GATTO par exemple, utilisé par le précieux TLIO, serait un bon instrument pour tel usage spécifique, et quelles en sont les limites pratiques et méthodologiques.

La présente contribution décrit les exigences philologiques pour l'étude des textes médiévaux assistée par ordinateur et les compétences techniques du programme. En outre, elle expose la problématique des décisions concrètes qu'implique la lemmatisation des textes anciens et les avan-

tages que celle-ci peut apporter du point de vue philologique aussi bien que linguistique.

Exigences philologiques

Toute élaboration de corpus, à des fins linguistiques ou d'édition, se fonde sur des faits philologiques préétablis. Par conséquent, ces faits déterminent la réflexion sur les outils de lemmatisation. Si l'analyse d'un corpus contemporain où les textes (journaux ou romans, par exemple) sont rédigés dans une langue standard bien décrite demande très peu de préparation — version unique, nombre réduit de signes, présentation linéaire, faible variation —, les textes médiévaux dépendent d'une tradition et d'une transmission extrêmement complexes.

La transcription, première étape d'une édition, est déjà une opération lourde et délicate. L'analyse électronique requiert la même qualité de transcription que dans le cas des éditions de référence sur papier. Les différentes stratégies éditoriales — éditions critiques basées sur un ou deux manuscrits de référence ou transcriptions (semi-) diplomatiques d'un ou de plusieurs manuscrits — peuvent certes se prévaloir d'un degré égal de qualité et être interprétées avec succès par des outils informatiques. Mais elles conduisent à des besoins et résultats informatiques différents suivant leurs qualités particulières : une édition critique contient déjà une information linguistique bien préparée et une interprétation du texte ; des transcriptions de manuscrits reflètent plus immédiatement l'authenticité des sources (cf. Glessgen et Lebsanft, 1997 : 10s.).

Une interrogation par ordinateur, basée sur des éditions critiques de bonne qualité, donnera d'excellents résultats pour l'étude du changement linguistique. Toutefois, le potentiel de quantification de l'informatique permet aussi

d'interroger à la fois plusieurs manuscrits parallèles transcrits minutieusement, ce qui serait presque impossible avec des méthodes traditionnelles. La variation entre différents manuscrits à l'intérieur d'une même tradition textuelle exige plus de réflexion ; mais il est d'ores et déjà possible de voir que l'informatique pourrait permettre la réalisation des postulats de la Nouvelle Philologie et baser l'analyse linguistique (et littéraire) sur l'évidence concrète des textes (cf. Oesterreicher 1997).

Pour la lexicologie de l'ancien français, la variation pourrait présenter un certain intérêt. La pratique de la glossographie française n'a jamais intégré les variantes textuelles contenues dans l'apparat critique — sans parler des variantes lexicales qui ne sont pas consignées dans l'édition (cf. Korfanty 1999). Toute la lexicologie aussi bien que la lexicographie de l'ancien français repose sur les formes plus ou moins normalisées des éditions critiques, en dépit du jugement des philologues qui sont bien conscients de ce problème. C'est la considération de toutes les formes des divers manuscrits d'un texte ou d'une tradition textuelle qui nous procurera toute l'information qu'on puisse connaître. Comme les textes sont répétitifs, nous n'aurons probablement pas de vraies surprises, mais certaines formes isolées pourront être identifiées et, pour ce qui est des formes les plus employées, la qualité de leur emploi sera mieux connu, en termes de fréquence, de syntagmatique et de répartition dans les sources[1].

1. Cf. Martin-Dietrich Glessgen, «L'élaboration philologique et l'étude lexicologique des Plus anciens documents linguistiques de la France à l'aide de l'informatique», dans Frédéric Duval (dir.), *Frédéric Godefroy. X^e colloque international sur le moyen français*, Metz, 12-14 juin 2002 (sous presse).

Même dans le cas d'une transcription de manuscrit étroitement fidèle, l'édition (électronique) devrait fournir des informations sur la segmentation et la fiabilité des mots ainsi que sur la structure du texte. Il est peu coûteux pour le transcripteur d'introduire apostrophes et signes pour la séparation ou la contraction des mots (par ex. *a-l'abé* pour *alabe*, ou *a-l'a_bé* pour *ala be*) : dans une représentation du texte contrôlée informatiquement, on peut facilement renoncer à ces indications et donner une vue authentique du document ; mais de cette façon, l'interrogation par ordinateur dispose en même temps d'une segmentation de mots cohérente, condition requise pour la plupart des recherches linguistiques. ; il est même possible de traiter immédiatement la question de la séparation et de la contraction des mots en vue d'étudier les habitudes d'un scribe.

Il importerait aussi de noter les *lapsus calami* et les bourdes des copistes médiévaux : *conite* n'est pas un mot français, alors que c'est le cas pour *comte*. Une étude des habitudes d'un scribe ou de sa conscience linguistique peut s'appuyer sur les indications de corrections qui figurent dans les notes. Mais le but d'une interrogation lexicologique, ce sont les lexèmes comme savoir intersubjectif et partagé.[2]

En bref, un corpus informatisé qui serve de base à la lemmatisation devrait être dépourvu d'erreurs formelles ; les corrections doivent être signalées, mais le travail lexicographique et grammatical doit s'effectuer à partir de formes linguistiquement cohérentes.

Le problème le plus délicat est celui de la structure ou segmentation du texte : il est possible de se fonder sur les lignes du manuscrit original, sur celles d'une transcription

2. Cette conviction a été forgée notamment dans des discussions avec Georges Kleiber (Strasbourg).

électronique (comme dans le corpus d'Amsterdam) ou sur celle d'une édition sur papier. Mais tout utilisateur ultérieur aura besoin de la version textuelle exacte sur laquelle se fonde l'interrogation ; en outre, les divers manuscrits d'une tradition textuelle unique auront des segmentations complètement différentes. Une segmentation qui corresponde au contenu des textes présente de sérieux avantages : elle est indépendante d'un antécédent particulier, contingent ; elle est facile à reproduire et à adapter pour divers manuscrits (au besoin par une référence double, une pour le texte et une pour le manuscrit : div n=3298, i=3187) ; et elle indique aussi dans quelle partie du texte on peut trouver les mots ou formes spécifiques.

Choisir une édition critique ou un manuscrit de référence approprié pour un texte, ou une tradition textuelle, donné exige une étude poussée aussi bien dans un contexte informatique qu'en philologie traditionnelle. Il faut accorder une attention particulière au choix des manuscrits : le grand investissement de temps nécessaire pour la transcription minutieuse d'un manuscrit permet seulement dans de rares cas la considération de tous les manuscrits d'un texte. Dans le cas des huit manuscrits du *Chevalier de la charrette*, présentés ici par Cinzia Pignatelli, le nombre des manuscrits est encore réduit et correspond à l'importance du texte. Mais pour les études textuelles tout comme pour l'informatique, ce serait sûrement une perte de temps que de se mettre à transcrire le nombre élevé de manuscrits (environ 600) de la *Divine Comédie* : nos énergies sont limitées par rapport au vaste champ des études de linguistique historique, et nous devons choisir où concentrer nos efforts.

Ce travail d'élaboration doit être accompagné d'une description diasystématique du texte et des manuscrits : il faut pour les deux une date et une localisation ; on doit

identifier aussi le genre de tradition textuelle et les modèles du texte, souvent latins. Ces éléments déterminent la forme linguistique du texte et doivent donc être intégrés dans l'interrogation linguistique.

Une étude basée sur des manuscrits originaux peut aussi tenter d'identifier la chancellerie ou le *scriptorium* où le manuscrit a été composé ou copié : ces lieux d'écriture constituent les entités sur lesquelles se fonde toute l'histoire linguistique du Moyen Âge, en termes de temps, d'espace et de tradition textuelle, mais aussi en termes de sociologie (cf. Gleßgen 2001). Cette partie de la description philologique ne peut se faire qu'après (ou en simultanéité avec) l'analyse linguistique ; mais il faut y viser dès le début.

Les données élaborées à partir des manuscrits doivent ensuite être encodées minutieusement de façon à garantir leur longévité. Ce besoin est un point crucial, qui a été longtemps et malheureusement sous-estimé dans les études textuelles assistées par ordinateur. Les langages d'encodage neutres comme XML, UNICODE et les propositions de la TEI peuvent maintenant procurer cette garantie. Néanmoins leur application n'est pas une mince opération : il faut étiqueter les données textuelles et la description diasystématique, justifier les listes d'étiquettes et instaurer une procédure de correction du corpus. Ces opérations peuvent être maîtrisées intellectuellement et techniquement, encore faut-il y prêter une grande attention, car elles exigent des compétences en informatique aussi bien qu'en philologie.

Les exigences philologiques sont, pour l'informatique, encore plus rigoureuses que pour des réalisations de type traditionnel. Les indications que nous avons brièvement exposées (interaction entre édition critique et manuscrits, choix des manuscrits, transcription d'une grande fidélité, toilettage du texte accompagné de justifications, segmenta-

tion du contenu, description diasystématique, encodage neutre) sont déterminées par l'attention portée parallèlement sur le respect de la matérialité du texte, sur les buts linguistiques et sur les fondements informatiques.

Le corpus de chartres que j'étudie dans le contexte des *Plus anciens documents linguistiques de la France* ne présente pas les mêmes problèmes que d'autres corpus comprenant des textes littéraires ou scientifiques. Comme les documents de la pratique juridique sont généralement des originaux, la question de l'interdépendance des textes et de l'écart entre un texte et une rédaction manuscrite ne se pose guère. Mais on y retrouve les autres caractéristiques de la variance médiévale. Nous avons déjà décrit notre corpus particulier et son élaboration à plusieurs occasions, nous ne nous répéterons pas ici[3] (cf. Glessgen 2001, s.p. a, s.p. b). Nous en donnerons seulement les traits essentiels : le corpus de base correspond aux 290 documents les plus anciens conservés dans les *Archives départementales de Meurthe-et-Moselle* à Nancy (1232-1265), édités très soigneusement par Michel Arnold sous forme dactylographiée. Travaillant en équipe, nous avons numérisé le texte, corrigé entièrement la transcription, intégré majuscules et signes de ponctuation originaux, introduit une segmentation de contenu et complété la description diasystématique. Le projet se place dans le contexte de l'édition et de l'étude des *Plus anciens documents linguistiques de la France*, sous la direction de Françoise Vielliard et Olivier Guyotjeannin.

3. Cf. Martin-Dietrich Glessgen, «Editorische, lexikologische und graphematische Erschließung altfranzösischer Urkundentexte mit Hilfe von TUSTEP. Stand der Arbeiten», dans Kurt Gärtner et Günter Holtus (dir.), *Drittes Trierer Urkundensprachekolloquium*, 20.-22. Juni 2001 (sous presse) et l'article cité dans la note 1.

Notre but est de préparer l'élaboration informatique et lexicologique de la collection.

Le texte encodé (*txt* - */txt*) et la description diasystématique du texte (*an* - */an*) se présentent de la façon suivante :

```
<gl>
<t type ="123"/>
<id>555550002</id>
<zitf>002</zitf>
<an>
<nom>002</nom>
<d>1234 (25 mars-31 d%/ecembre) ou 1235 (1#'e#'r janvier- 24
mars)</d>
<d0></d0>
<type>charte : acensement de terres</type>
<r>L'abb%/e et le chapitre de Salival acensent %\a Wirrion
et Houillon treize journaux de terre au finage de Juvelize
contre un cens de treize deniers et deux h%/emines de grain;
les conditions de l'acensement sont tr%\es contraignantes
pour les paysans.</r>
<aut>non annonc%/e</aut>
<disp>abbaye de Salival</disp>
<s>disposant</s>
<b>disposant [la r%/edaction de la charte avantage surtout
le chapitre]</b>
<act>Wirrion et Houillon, paysans de Juvelize</act>
<rd>scriptorium de l'abbaye de Salival [les paysans ne
pouvaient pas disposer d'un scribe]</rd>
<f>Parchemin jadis scell%/e sur simple queue; 58x141</f>
<l>AD MM H 1244, fonds de l'abbaye de Salival</l>
</an>

<txt>
<pub><div n=1><maj>C</maj>onue chose soit a-toz</div> </pub>
<exp><div n=2> q<abr>ue</abr> li abes <abr>et</abr> li
chapitles de Salinvas /. at laissi%/e a Wirion <zw/>
<abr>et</abr> Huillon, les dous freres de Gev<abr>er</
abr>lise, les anfanz Bertran Bacheler,</div>
<div n=3>/.XIII/. jor<zwt/>nas de t<abr>er</abr>re treisse
/. en la fin de Gev<abr>er</abr>lise /. <abr>et</abr> a lor
oirs /.</div>
<div n=4> p<abr>ar</abr>mi /.XIII/. d<abr>eniers</abr> de
cens /. <abr>et</abr> <zw/> /.II/. himas de blef /. l'un
d'avoine /. l'autre de froment /.</div>
<div n=5> <abr>et</abr> s'il ne paievent a jor <zw/> nomei a
la feste sent Remi /. a Giv<abr>er</abr>lise, en la maison
de Salinvas <ful>L'abbaye poss%\ede donc une maison %\a
Juvelize.</ful> /. q<abr>u</abr>e l'on se tan <zwt/>roit a
la terre /. <abr>et</abr> ce q<abr>ue</abr> sus averoit /.</
div>
<par><div n=6> (...) </par> </exp>
<par><cor> <div n=10> <maj>C</maj>i at mis li abes <abr>et</
abr> li covenz de Salinvas son sael /. en tesmoig<zwt/>nage
de verit%/e /.</div></par> </cor>
```

```
<dat> <div n=11> l'an q<abr>ue</abr> li miliaires corroit
p<abr>ar </abr> /.M/. <abr>et</abr> CC/. <abr>et</abr>
XXXIIII/. anz /.</div></par> </dat>
</txt>
</gl>
```

Comme le texte est en cours d'élaboration, la liste d'étiquettes comporte encore certaines singularités ; il faudra l'adapter aux règles propres à xml et aux propositions de la TEI, mais la transformation automatique est déjà programmée.[4]

Voici une version imprimée de ces données, générée automatiquement :

002

1234 (25 mars-31 décembre) ou 1235 (1er janvier-24 mars)

Type de document : charte : acensement de terres
Objet : *L'abbé et le chapitre de Salival acensent à Wirrion et Houillon treize journaux de terre au finage de Juvelize contre un cens de treize deniers et deux hémines de grain ; les conditions de l'acencement sont très contraignantes pour les paysans.*

Auteur : non annoncé
Disposant : abbaye de Salival
Sceau : disposant
Bénéficiaire : disposant [la rédaction de la charte avantage surtout le chapitre]
Autres acteurs : Wirrion et Houillon, paysans de Juvelize

4. L'application des balises est contrôlée par une procédure programmée sous XMLSPY.

Rédacteur : scriptorium de l'abbaye de Salival [les paysans ne pouvaient pas disposer d'un scribe]

Parchemin jadis scellé sur simple queue ; 58 x 141
AD MM H 1244, fonds de l'abbaye de Salival

1 Conue chose soit a-toz **2** q*ue* li abes *et* li chapitles de
5 Salinvas · at laissé a Wirion / *et* Huillon, les dous freres de Gev*er*lise, les anfanz Bertran Bacheler, **3** ·XIII· jor/nas de t*er*re treisse · en la fin de Gev*er*lise · *et* a lor oirs · **4** p*ar*mi ·XIII· d*eniers* de cens · *et* / ·II· himas de blef · l'un d'avoine · l'autre de froment · **5** *et* s'il ne paievent a jor // nomei a la feste sent Remi · a Giv*er*lise, en la maison de Salinvas[1] · q*ue* l'on se tan/roit a la terre · *et* ce q*ue* sus averoit ·
6 [...]
10 Ci at mis li abes et li convenz de Salinvas son sael · en tesmoig/nage de verité · **11** l'an q*ue* li miliaires corroit p*ar* ·M· *et* CC· *et* XXXIIII· anz ·

[1] L'abbaye possède donc une maison à Juvelize.

Les principes informatiques que nous appliquons à notre corpus sont très semblables à ceux suivis par le *Dictionnaire du Moyen Français*, par Christiane Marchello-Nizia ou par Pierre Kunstmann ; aussi les corpus peuvent-ils être facilement comparés.

On peut naturellement se poser la question de savoir si tous ces efforts vont donner des résultats. En d'autres mots : notre vision de l'histoire de la langue changera-t-elle si elle se fonde un jour sur l'interrogation de corpus électroniques constitués suivant les critères exposés? Une autre question est celle du montant de travail requis pour la préparation d'un tel corpus.

Je ne peux, honnêtement, répondre à ces questions pour le moment. Le corpus du TLIO, qu'on peut considérer, dans le domaine roman, comme le corpus de référence préparé avec le plus d'attention aux questions philologiques, a

déjà transformé les études lexicologiques portant sur l'italien ancien. Même s'il ne va que jusqu'en 1375, le corpus a créé une nouvelle base pour la rédaction de la partie médiévale du *Lessico Etimologico Italiano* : nouvelles datations, nouveaux sens, nouveaux contextes et — ce qui est encore plus important — une meilleure référence pour la définition de sens déjà connus. Cependant, le vocabulaire de l'ancien français est bien mieux connu que celui de l'ancien italien et une étude fondée sur un quelconque corpus ne parviendra certainement pas aux mêmes résultats. Les nouvelles perspectives ouvertes par l'informatique et l'établissement de corpus fiables portent sur des questions d'usage linguistique (fréquence, contextes, tradition textuelle) et sur le processus d'élaboration d'une langue préstandardisée. Dans ces domaines, en revanche, on peut être sûr qu'une approche à partir des manuscrits conduira à des résultats nouveaux et intéressants.

Mais intégrer dans le domaine de l'informatique un héritage philologique élaboré en critique textuelle durant 150 ans constitue une tâche énorme et lourde, qu'il faut entreprendre avant (ou parallèlement à) la transcription de manuscrits, connus ou inédits. Dans le processus de numérisation de différents types de textes anciens, il est important d'indiquer systématiquement les caractéristiques philologiques des textes respectifs : édition critique, basée sur un ou plusieurs manuscrits, transcription semi-diplomatique, qualité du travail. Dans le projet qui nous unit à Pierre Kunstmann et Achim Stein, nous adaptons le *Corpus d'Amsterdam* — préparé minutieusement par Antonij Dees et Piet van Reenen — à ces exigences fondamentales ; nous nous préoccupons, en particulier, d'évaluer la qualité des textes. Il est clair que le travail lexical, ou grammatical, sur

un corpus d'une telle importance sera utile même si tous les critères ne sont pas remplis.

L'interprétation linguistique laissera tomber la majeure partie des étiquettes introduites pour certaines raisons philologiques ; mais il reste important de savoir sur quel type de données se fonde l'analyse. Finalement — et ceci explique ces longs préliminaires — les outils d'interprétation doivent respecter les compétences philologiques, et non l'inverse, comme le souligne Ineke Hardy dans le présent volume.

Fonctionnalités du lemmatiseur

La fonctionnalité de base d'un lemmatiseur est de relier diverses formes (*types*) et occurrences (*tokens*) à une entité distincte. Un lemme comme *abbé* subsume un grand nombre de formes graphiques (*abey, abbeis*, etc.) et se trouve relié à leur apparition concrète dans un corpus textuel. Une telle opération s'apparente à la création d'autres entités d'intérêt linguistique comme des éléments morphologiques, syntaxiques ou graphématiques ; dans ces cas, des formes partageant certaines particularités avec d'autres formes sont unies et cette unification constitue la base d'interrogations linguistiques subséquentes. Le lemmatiseur que nous mettons au point peut aussi servir à des interrogations d'ordre onomastique, graphématique ou grammatical. Dans tous les cas, l'outil informatique doit permettre une organisation de formes de manière rapide et contrôlée.[5]

5. La programmation se fait dans le cadre d'une collaboration entre Matthias Kopp (Tübingen) — pour le côté informatique — et l'auteur de ces lignes — pour le côté linguistique et philologique. Nous utilisons un langage script, *Tustep*, qui n'est pas très diffusé (par rapport à *Perl*, par ex.), en particulier en dehors de la communauté scientifique de langue allemande, mais qui est très

Le meilleur principe directeur pour assembler les formes sous un lemme identique nous paraît être l'identité étymologique : nous considérons comme formes d'un seul lemme celles qui appartiennent au même paradigme morphologique et qui ont la même origine. La variation sémantique à l'intérieur d'un tel lemme correspond aux principes cognitifs de transformation (métonymie/métaphore) et peut être structurée en tant que telle.

La procédure que nous avons adoptée est, au début, très facile et traditionnelle : le lemmatiseur produit un index KWIC (*key words in context*) avec toutes les occurrences contenues dans le corpus et donne le contexte d'une ligne ainsi qu'une référence marquant la position de cette occurrence dans le texte. Une première procédure automatique permet de séparer mots et noms propres : les formes commençant par une majuscule introduite par l'éditeur (et non les majuscules d'époque) sont rassemblées dans un index des noms propres (à l'exception des formes situées en début de phrase) ; chaque forme commençant par une minuscule apparaît dans l'index lexical.[6] Les deux index peuvent être construits de la même façon, mais nous nous concentrerons ici sur les lexèmes.

Une seconde intervention semi-automatique porte sur les cas d'« équivalences » graphiques qui dépendent du système graphématique du langage étudié. Les textes anciens de langue romane — ou germanique — présentent un nombre élevé d'équivalences graphématiques quasi absolues ; dans notre corpus, un double *b* n'est jamais placé en forte opposition graphématique avec un simple *b* ; tout

bien conçu, avec des fonctions qui se prêtent excellemment à l'étude philologique.

6. Les noms propres en début de phrase doivent être étiquetés.

mot écrit, pour des raisons étymologiques, avec *bb* peut être écrit avec *b* (par ex. *abbaie* - *abaie* < ABBATIA), même s'il n'existe pas de réduplications non étymologiques dans le même corpus. Des équivalences du même type arrivent entre d'autres doubles consonnes et les consonnes simples qui leur correspondent (*cc* - *c*, *dd* - *d*, etc.). Les doubles voyelles sont moins utilisées, mais on les rencontre aussi dans le corpus, sans valeur distinctive particulière (*aa* - *a*, *ee* -*e*, etc.). Cette variance graphique n'est pas seulement typique de l'ancien français ; elle caractérise aussi habituellement les époques et les langues (tout comme les individus) dont l'usage scripturaire n'est pas pleinement maîtrisé (cf. Ernst/Wolf 2002).

On trouve d'autres équivalences utiles pour relier des formes d'origine commune, même si elles répondent à des fonctions différentes en termes d'histoire de la langue. Nous avons relevé dans notre corpus :

- des homophones : *en / an*, *y / i*, *-y / -s*, *(n)gni / ngn / gn* ;
- des groupes consonantiques latinisés : *cq / q*, *ct / t* ;
- des redondances graphiques : *k / qu*, *-x / -us* ;
- des variations régionales d'ordre graphématique ou phonématique : *-ei / -é*, *-eir / -er*, *np / mp*, *w / g*.

La validité de telles équivalences est évidemment limitée dans le temps, dans l'espace et même dans les genres littéraires. Dans un corpus homogène comme nos chartes lorraines, on peut définir et appliquer de telles équivalences sans problème particulier. Mais en définir pour de larges corpus exige l'intégration de faits diasystématiques : si l'on peut appliquer *bb=b* à toutes les formes et à tous les textes d'avant 1540 dans tout l'espace du français, *-ei / -é* ne concernerait que les formes d'avant 1450 dans la partie Nord-

Est de la *langue d'oïl*. Une telle application rend nécessaire la définition diasystématique des textes.[7] L'intérêt linguistique de la définition d'équivalences est évident : elles représentent une première nomenclature pour l'analyse de la variance graphématique ; toute équivalence peut faire l'objet d'une interrogation en termes de fréquence, de chronologie et même de diasystème. Ces équivalences sont aussi très utiles pour la lemmatisation : une application de quelque 40 équivalences sur notre corpus a réduit le nombre d'entrées de moitié ; un très petit nombre de formes sont réunies par erreur et doivent être séparées. Par exemple, la forme générée *abé* «abbé» regroupe les formes graphiques *abé, abbé, abei, abbei* et *abey* ; la forme *abés* (sujet singulier et régime pluriel) inclut *abés, abbés* et *abey* ; la forme non attestée *abet*, construite automatiquement, donne *abeit* et *abbeit* ; les formes isolées *ebbé* et *ebbeit* génèrent des entités simples. Finalement, le lemme a.fr. *abbé* correspond, dans notre index KWIC préstructuré, à cinq entrées au lieu de douze.

L'index KWIC se présente avec une interface tripartite :

1) liste de toutes les formes de base construites automatiquement ;
2) en cliquant sur une forme, l'utilisateur peut trouver toutes les graphies que regroupe cette forme ; chaque occurrence apparaît sur une seule ligne avec un bref contexte et l'indication de la référence au texte ;
3) un clic près d'une occurrence spécifique fait apparaître un contexte plus large de cette occurrence, lequel peut être encore élargi dans une nouvelle fenêtre.

7. Il est à noter que nous n'avons pas encore programmé l'application d'une équivalence dépendant de critères diasystématiques.

La structure tripartite correspond au travail traditionnel du glossographe qui lemmatise un texte : il peut repérer chaque forme et décider, en fonction de son savoir linguistique, si les formes reliées appartiennent ou non à un lemme unique. Le clic sur une occurrence (2) permet de sélectionner une forme ; une fois que toutes les occurrences à regrouper sont sélectionnées, une forme lemmatique peut leur être attribuée. Pour cette opération est utilisée la liste des lemmes déjà introduits.

L'étape suivante (non pour l'utilisateur, mais pour le programme) est un étiquetage automatique des formes sélectionnées dans le texte authentique : chaque forme reçoit une nouvelle étiquette placée dans le texte et contenant un numéro de mot spécifique (par ex., *Conue chose soit a-toz que li <wn n =328 lex=abbé> abes </wn> et li chapitles de Salinvas...*).[8] L'élaboration ultérieure des formes unies sous un lemme se fondera sur les résultats de cette procédure semi-automatique. Sur la base du texte authentique, lemmes et formes graphiques sont générés immédiatement et peuvent être intégrés dans une structure de base de données.

Un avantage de ce principe pour l'éditeur de texte est qu'il peut encore introduire des changements et des corrections dans le même texte sans perdre les informations déjà introduites par la lemmatisation. Ce qui est très important pour le travail concret, puisque l'élaboration d'un glossaire — l'un des buts de la lemmatisation — conduit normalement à des corrections dans le texte de départ.

8. En même temps les nouvelles indications sont placées dans un index séparé (<wn n=328> <lex>abbé </wn>) ; ce double enregistrement répond à l'architecture actuelle du programme, mais n'est pas nécessaire.

Les procédures de base de notre lemmatiseur peuvent être adaptées à des fins diverses. Une fonction supplémentaire banale est la restriction à certaines tranches de fréquence. La lemmatisation peut s'appliquer à une tranche sélectionnée : les mots de fréquence élevée ont plutôt des fonctions grammaticales et demandent un traitement très différent de celui réservé aux mots de basse fréquence, qui en principe ont un contenu lexical spécifique.

Nous n'avons pas encore établi d'interface entre cette procédure de lemmatisation et d'autres sources, en particulier des étiqueteurs morphologiques — comme le *TreeTagger* qu'a élaboré Achim Stein et qu'il décrit dans le présent volume — ou des vocabulaires de référence. La complémentarité des deux procédures avec notre lemmatiseur est évidente et leur intégration fait partie de nos projets à venir.

Les fonctions actuelles de notre lemmatiseur présentent différentes qualités : il reproduit vraiment le travail traditionnel et philologique tout en lemmatisant, mais il accroît sa puissance de façon remarquable ; il maintient une relation vivante entre l'édition du texte et la base lemmatisée, ce qui permet d'apporter ultérieurement des corrections au texte ; il peut être réglé par l'utilisateur pour les besoins de n'importe quel texte de langue romane, même pour des sources orales qui présentent, pour la variance, des caractéristiques analogues à celles des textes historiques (cf. Pusch/Raible 2002) ; il fonctionne sans vocabulaire de référence ou étiquetage morphologiques préalable, mais reste ouvert à l'intégration de procédures complémentaires.

Actuellement, je n'ai pas encore pu comparer en détails les qualités de cet instrument avec celles d'autres outils, commerciaux ou pas ; mais j'ai presque atteint mon but de trouver un terme de comparaison pour déterminer les capacités d'un lemmatiseur.

Développements à venir et perspectives

Les formes et occurrences appartenant à un lemme une fois identifiées, le vrai travail de lexicologie historique en est encore à ses débuts : il faut structurer les formes suivant leurs traits morphologiques et sémantiques. Cette étape, cruciale dans notre procédure, est en cours. Le meilleur système pour la base lexicologique que nous voulons établir est, à notre avis, le principe du DÉAF : ce dictionnaire exemplaire présente d'abord tous les exemples dans un ordre morphologique et formel ; ce faisant, il fait apparaître de façon claire la distribution diachronique et diasystématique des différentes formes ; une application informatique peut ajouter des informations concernant la fréquence. Dans une seconde section, le DÉAF organise les différents sens d'une unité lexicale, en incluant les collocations et les phraséologies. Les deux séries de «formes» et de «sens» seront accompagnées de commentaires plus ou moins développés.

Mais il reste encore à résoudre une large série de questions. La lemmatisation concrète d'un corpus est extrêmement variable comme l'a montré S. Korfanty (1999). Nous avons rejeté la conception très stricte du lemme (un sens, un cadre de valence) et retenu le traitement des formes polysémiques fondé sur l'étymologie, incluant leurs différents cadres de valence et leurs divers contextes syntagmatiques. Mais la variance élevée de l'ancienne langue crée des problèmes épineux ; la même racine présente des variantes importantes d'une région ou d'une période à une autre, même dans un corpus étroit comme le nôtre : faut-il traiter les trois formes *boter*, *aboter* et *abotir* comme un, deux ou trois lemmes? Le changement de catégorie grammaticale *mangier* (verbe — nom) doit-il être placé sous un seul lemme? Même la dérivation pose des problèmes : les cadres sémantiques et pragmatiques de formes provenant de la

même base mais suffixées différemment sont souvent très proches. Il est facile de relier des formes proches, mais il convient d'établir un équilibre pour les lexèmes d'une même origine entre leur intégration sous une même entrée et leur séparation.

Suivre jusqu'au bout cette lourde procédure répond à plusieurs finalités :

1) En premier lieu, cela permettra l'établissement d'un glossaire bien construit ; cette finalité est utile pour tout texte ancien, et c'est une des raisons initiales de notre programmation : le lemmatiseur devrait servir aux spécialistes de critique textuelle travaillant sur des sources très divergentes sans autre outil ou préparation informatique. Dans le cas des *Plus anciens documents linguistiques de la France,* il y a l'avantage supplémentaire de pouvoir dériver différents glossaires à partir d'une base lexicologique unique : l'auteur de tout nouveau volume de la collection peut utiliser la structure et les définitions déjà établies, il lui suffit d'ajouter ses propres formes, sens et lemmes. À la fin de l'entreprise (dans dix ou quinze ans), on aura constitué un nouveau vocabulaire fiable pour la langue des documents.

2) Deuxièmement, la procédure peut aussi servir aux bases textuelles déjà existantes ou servir à relier bases textuelles et bases lexicales. Elle peut contribuer également à l'établissement d'une relation plus directe entre corpus de textes et dictionnaires historiques, ce qui est actuellement une des perspectives les plus importantes en lexicologie et lexicographie historiques.

3) Finalement, les formes relevant d'un lemme peuvent être interprétées en termes d'histoire de la langue : le degré de standardisation d'une langue, par exemple, peut être mesuré par le degré de variation à l'intérieur des entités lemmatisées ; ou la complétude d'un paradigme verbal pourrait indiquer son degré d'élaboration. Enfin le domaine le plus passionnant est peut-être celui du changement linguistique à l'intérieur des filiations étymologiques qu'apportent les lemmes : les changements d'ordre formel, sémantique et syntagmatique ou la substitution de mots peuvent s'étudier sur cette base.

Bibliographie

GLESSGEN, Martin-Dietrich (2001), «Das altfranzösische Geschäftsschrifttum in Oberlothringen : Quellenlage und Deutungsansätze», dans Kurt GÄRTNER, Günter HOLTUS, Andrea RAPP et Harald VÖLKER (dir.), *Skripta, Schreiblandschaften und Standardisierungstendenzen* (Beiträge zum Kolloquium vom 16. bis 18. September 1998 in Trier), Trier, THF, p. 257-294.

GLESSGEN, Martin-Dietrich, et Franz LEBSANFT (1997), «Von alter und neuer Philologie oder : Neuer Streit über Prinzipien und Praxis der Textkritik», dans Martin-Dietrich GLESSGEN et Franz LEBSANFT (dir.), *Alte und neue Philologie*, Beihefte zur *editio* 7, Tübingen, Niemeyer, p. 1-14.

KORFANTY, Sylvie (1999), *Lexicographie et glossographie du français du XVIe siècle*, Thèse dactylographiée, UMB Strasbourg.

OESTERREICHER, Wulf (1997), «Sprachtheoretische Aspekte von Text-philologie und Editionstechnik», dans Martin-Dietrich GLESSGEN et Franz LEBSANFT (dir.), *Alte und neue Philologie* (Beihefte zur *editio* 7), Tübingen, Niemeyer, p. 111-126.