Bases de données textuelles et lexicographie historique: l'exemple des Plus anciens documents linguistiques de la France

Glessgen, M D

Martin-D. Gleßgen

Bases de données textuelles et lexicographie historique : l'exemple des *Plus anciens documents linguistiques de la France*[1]

1. La base empirique

Les réflexions suivantes reposent sur le projet concret de l'édition et de l'analyse linguistique des *Plus anciens documents linguistiques de la France*. La reprise de ce projet, après la mort de Jacques Monfrin, par Françoise Vielliard, Olivier Guyotjeannin et l'auteur de ces lignes s'est faite d'emblée sous les auspices de la philologie informatique. Les textes à éditer étaient destinés à être saisis sous une forme permettant parallèlement une édition papier et une édition électronique ainsi qu'un traitement linguistique assisté par ordinateur.

Ce n'est pas le lieu pour présenter en détail le projet des *Plus anciens documents* ; voici seulement quelques données très générales : le projet entend répertorier et éditer tous les textes documentaires en français et en occitan dans la première période de leur mise à l'écrit. Le relevé est organisé par lieu de conservation des documents originaux, à commencer par les départements de la France. L'étude s'arrête en principe en 1270 (date retenue par Jacques Monfrin) ; mais puisqu'il s'agit de documenter la mise à l'écrit dans tout l'espace galloroman, la date peut être repoussée pour des régions où l'écrit vernaculaire apparaît plus tardivement (p. ex. 1290 pour le Jura), ou elle peut aussi être avancée en cas de documentation très dense (p. ex. 1265 pour la Meurthe-et-Moselle). Actuellement, dix-sept départements français de la langue d'oïl sont en chantier, essentiellement dans la moitié nord-est du domaine d'oïl ; parallèlement, des équipes suisse et belge travaillent sur les documents conservés en Suisse Romande et en Belgique ; le redémarrage de série occitane est en préparation.

Les observations méthodologiques suivantes reposent sur cette entreprise mais s'inscrivent en même temps dans une réflexion plus générale sur la relation entre la philologie informatique et l'historiographie linguistique et, de manière plus spécifique, entre les bases de données textuelles et la lexicographie historique.

---

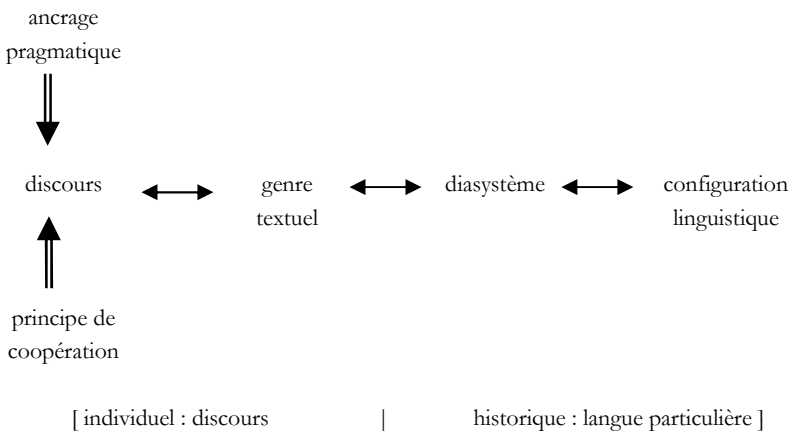
[1] L'aperçu présenté ici résume les éléments d'un atelier informatique présenté à Lecce ; il s'agissait de mettre en relief précisément notre recherche et sa méthodologie concrète, ce qui explique les quelques renvois bibliographiques à celle-ci. Ce texte est identique à celui publié dans la section de *De la philologie aux nouveaux médias* du XXIV[e] Congrès de la *Société de Linguistique et de Philologie Romanes* (Aberystwyth, 2004, sous presse), qui s'adresse à un public différent.

2. Philologie (informatique) et historiographie linguistique

La relation entre la philologie et l'historiographie linguistique est caractérisée par une opposition sous-jacente entre le “texte” et la “langue”. L'historiographie linguistique étudie l'évolution de la langue à travers le temps, le diasystème et les genres textuels. La philologie au contraire s'interroge sur le sens d'un texte ou discours individuel à un moment donné et sur les signes linguistiques qui produisent ce sens.

Naturellement, il existe une interaction forte entre les deux : la philologie repose sur le savoir de l'histoire linguistique et l'historiographie de la langue utilise les études philologiques. Prenons l'exemple saillant des premières attestations : tout philologue utilisera le FEW ou le LEI comme références pour identifier les premières ou dernières attestations contenues dans un texte qu'il édite mais, par cette même opération, il modifie la référence historiographique du mot en question.

Or, la philologie informatique peut réduire l'écart entre la “langue” et le “texte”. Elle repose sur des masses de données textuelles importantes qui créent un ensemble intermédiaire entre le texte individuel et le diasystème de la langue. L'étude d'un ensemble textuel important permet par la quantification de relever des variantes particulières et des usages généraux et fait disparaître les éléments événementiels et contingents, inhérents aux textes individuels. Notamment, la philologie informatique met en relief la dimension des genres textuels dans leur tradition historique, ce qui crée un ancrage sûr. Il faut supposer que les genres textuels occupent une position bien identifiable entre le discours individuel et le diasystème et la configuration linguistiques ; cf. le schéma suivant (d'après Gleßgen 2005) :



Par les regroupements des genres textuels que permet la philologie informatique, l'historiographie linguistique et la lexicologie historique peuvent s'appuyer directement sur des témoins linguistiques concrets.

Les questions philologiques traditionnelles restent néanmoins entières ou se posent même avec plus de rigueur : il faut toujours considérer le type de sources (manuscrit unique *vs* tradition complexe) et définir le type d'édition (diplomatique, critique, synoptique) ainsi que le traitement des variantes. Même si l'édition électronique offre des solutions beaucoup

plus flexibles que l'édition papier, l'encodage des données exige une attention considérable, au moins dans un premier temps. Pour garantir la longévité des données textuelles – exigence qui ouvre de nouvelles perspectives mais qui est en même temps contraignante – il faut faire appel à un encodage particulier (type XML, support de UNICODE, choix des balises orienté d'après les propositions de la TEI [*Text Encoding Initiative*]) ; par ailleurs, le choix des outils de traitement et des programmes doit viser à son tour à la longévité, ce qui plaide en faveur d'outils de type *Open Source* comme *Linux* et *Open Office*.

De même, les questions de localisation et de datation gardent toute leur importance. Il s'agit de dater les textes et les manuscrits, de définir leur appartenance à un genre textuel donné, de les placer dans l'espace, d'identifier leur utilisation par un groupe social défini et, idéalement, de relever la chancellerie ou le scriptorium qui a produit le texte en question. Tous ces paramètres du diasystème doivent être définis au préalable pour permettre l'étude linguistique du corpus informatique.

Enfin, les exigences du commentaire linguistique de base restent entières : il faut préparer des glossaires lexical et onomastique et décrire les systèmes graphématique et morphologique. Ces éléments peuvent être utilisés dans une analyse linguistique ultérieure, en identifiant des entités définies (des lexèmes individuels, des champs sémantiques, des variantes graphématiques, des grammèmes particuliers) et en les soumettant à la quantification selon les paramètres du diasystème. Les analyses syntaxique et textuelle se placent, par leur complexité, à un niveau d'étude supérieur.

3. Édition et encodage électronique dans les *Plus anciens documents*

L'application de ces principes méthodologiques pour le traitement des *Plus anciens documents linguistiques de la France* montre les difficultés pratiques que soulèvent les différents niveaux d'étude ou d'analyse.

La présentation de l'édition imprimée reste relativement traditionnelle ; voici l'exemple d'une charte abrégée du corpus de la Meurthe-et-Moselle (cf. déjà Gleßgen 2003) :

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

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 Salinvas · at laissé a Wirion / *et* Huillon, les dous freres de Gev*er*lise, les anfanz Bertran Bacheler, **3** ·XIII· jor/nas de terre 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[2] · 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 ·

L'exemple illustre certains des principes d'éditions appliqués dans les *Plus anciens documents*, notamment pour les éléments de structure :
- les mots sont séparés selon l'usage moderne, mais avec l'indication univoque de l'intervention éditoriale : à partir des formes *a-toz* ou *l'un* l'ordinateur peut reconstituer la forme du manuscrit (*atoz*, *lun*) ou proposer une lecture très interprétative (*a toz*)[3] ;
- l'ajout des apostrophes et des accents suit ce même modèle : elles appartiennent toujours à l'éditeur et peuvent être supprimées dans une perspective diplomatique ;
- les majuscules du manuscrit sont rendues (dans la mesure où elles sont identifiables) en gras, les abréviations en italique, ce qui correspond à un encodage explicite (cf. *infra* : les balises <maj> </maj> et <abr> </abr>) ; toutes les autres majuscules sont introduites par l'éditeur : là encore, le fichier informatique permet de reconnaître autant la forme manuscrite que les interventions éditoriales ;
- les ponctuations médiévale et moderne sont distinguées ;
- le texte est segmenté de manière univoque pour fournir des références sûres (ici par <div n="1", cf. *infra* ; le système prévoit néanmoins des cas de figures plus complexes) ;
- les *lapsus calami* sont corrigés (tout en les indiquant), ce qui permet à l'analyse linguistique de reposer sur des formes de langue et non pas sur des erreurs d'écriture (p.ex. *comte* et non pas, dans une copie, *conite*).

L'innovation de cette édition repose surtout sur l'encodage informatique qui permet des interprétations différentes (diplomatique ou interprétative) selon les exigences du lecteur. Voici l'exemple de l'encodage, légèrement simplifié, pour notre charte :

```
<txt>
<pub.s/>
<div n="1"> <maj>C</maj>onue chose soit a-toz</div>
```

[2] L'abbaye possède donc une maison à Juvelize.
[3] Un autre exemple : la forme dans le manuscrit *alabe* sera transcrite *a-l'abé*, la forme manuscrite *ala be* comme *a-l'a_bé*; dans les deux cas, autant la lecture diplomatique que la lecture interprétative peuvent être dérivées automatiquement par un programme de transformation.

```
<pub.e/>

<exp.s/>
<div n="2"> q<abr>ue</abr> li abes <abr>et</abr> li chapitles de Salinvas /. at laissié 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> (…)
<par/>
<div n="6"> <maj>S</maj>i est ensi devisee /. q'au Tramble en
<zw/> at /.IIII/. jornas,//. un p<abr>ar</abr> lui <ful>Probablement <abr>Wirion</abr>, le
premier frère nommé dans le texte.</ful> /. <abr>et</abr> /.III/. ensemble;//.</div> (…)
<exp.e/>
<cor.s/>
<par/>
<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 verité,//. </div>
<cor.e/>
<dat.s/>
<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>
<dat.e/>
</txt>
```

L'encodage de type xml suit les principes évoqués en haut. Le texte est contenu entre les balises <txt> </txt> et structuré par les divisions sémantiques (<div etc.>) ainsi que par les parties des chartes (<pub[licatio]/>, <exp[ositio]/>, <cor[roboratio]/>, <dat[atio]/>) et des paragraphes (<par/>)[4]. Les différentes balises (cf. Gleßgen 2003) sont répertoriées dans un schéma. Après la saisie de la charte, il est donc possible, à l'aide d'un éditeur-xml, de vérifier si les balises correspondent au schéma et sont correctement posées[5]. Cette opération est contraignante mais elle garantit en même temps une qualité de saisie très appréciable.

Le fichier textuel peut être présenté sous différentes vues, grâce à un programme d'interprétation : il peut apparaître sous la forme proposée en haut mais aussi sous une forme entièrement diplomatique ou plus interprétative. Lors d'une présentation web, l'utilisateur peut même choisir, à partir du même fichier informatique, quelle vue il préfère. Ajoutons que la présentation web prévoit aussi les photographies des chartes.

Il est évident que l'encodage du fichier en question demande un effort plus important, au moins au début, que la saisie d'un texte par un programme comme Word. Mais les difficul-

---

[4] Puisque l'encodage xml n'admet pas de balises croisées, nous avons introduit pour les parties de la charte et les paragraphes des paires de *milestones* (<exp.e/> - <exp.f/>), ce qui permet de contourner l'obligation de placer toute paire de balises à l'intérieur d'autres paires de balises ; les deux éléments des *milestones* peuvent paraître dans deux divisions différentes.

[5] Nous utilisons l'éditeur *xml-spxy* dont la version 'home' est gratuite et facilement téléchargeable.

tés se posent surtout au début ; dans la durée, les avantages de l'homogénéité de l'encodage et de la longévité des données gagnent en importance ; par ailleurs, cet encodage permet non seulement des éditions variables mais aussi des interrogations linguistiques sur des bases limpides. L'élément charnière pour l'analyse linguistique est fourni par un lien immédiat entre les données textuelles et leur description diasystématique. De cette manière, toute forme linguistique est placée à tout moment dans son contexte d'usage.

4. Les paramètres du diasystème

Tout texte édité dans le cadre des *Plus anciens documents* est accompagné d'une description diasystématique détaillée. Les quatre paramètres qui me paraissent universellement applicables sont les suivants : le temps (encodé par <d>), le genre textuel (<type>), l'espace (<loc>) et le contexte social (<soc>) ; s'ajoute comme cinquième paramètre le rédacteur d'un document (<rd>) qui introduit une dimension individuelle ou idéosyncratique de celui qui a produit concrètement un texte. Voici l'encodage correspondant à la description diasystématique et philologique de notre exemple :

```
<id>555550002</id>
<zitf>002</zitf>

<an>
<nom>002</nom>
<d>1234 (25 mars-31 décembre) ou 1235 (1er janvier- 24 mars)</d>
<d0>1234/09/25</d0>
<type>charte: acensement de terres</type>
<loc>Lorraine ducale</loc>
<loc0>-</loc0>
<soc>Clergé régulier</soc>
<soc0>-</soc0>
<r>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.</r>
<aut>non annoncé</aut>
<disp>abbaye de Salival</disp>
<s>disposant</s>
<b>disposant</b>
<act>Wirrion et Houillon, paysans de Juvelize</act>
<rd>scriptorium de l'abbaye de Salival [rdp: néant; com: soustraitance?]</rd>
<rd0>AbbSalival</rd0>
<f>Parchemin jadis scellé sur simple queue; 58x141</f>
<l>AD MM H 1244, fonds de l'abbaye de Salival</l>
<ec>semi-onciale archaïque, frustre, statique, très lisible;
<form>s-</form> long systématique </ec>
```

<met>latinisme
<form>chapitle</form> (2)</met> </an>

Ce tableau analytique (<an> </an>) contient aussi des informations matérielles (lieu de conservation, éléments de description externe et linguistique du texte) ainsi qu'un résumé du contenu du document (<r[egeste]>). Tout texte (<txt>) est accompagné d'un tel tableau ainsi que d'une clé d'identification (<id> </id>; l'ensemble est balisé par <gl> </gl>).

Parmi les paramètres du diasystème, la date représente la seule donnée objective, puisque les chartes sont datées et, pour la plupart, originales. Mais dans le cas de copies, habituelles pour l'écrit non-documentaire, même la date est une donnée qui doit être déduite à partir d'éléments extérieurs.

Le genre textuel, déterminant pour la forme linguistique du document, ne peut être défini que sur la base de la structure textuelle (est-ce qu'il existe des structures différentes pour une vente et pour un acensement ? si oui, il faut distinguer deux sous-genres, sinon, il faut les réunir dans une catégorie supérieure).

L'espace autant que l'appartenance à un groupe social sont, dans les chartes, des éléments émergentes qui doivent être déduites de l'identification du rédacteur. Celle-ci repose à son tour sur une analyse externe (écriture <ec>, transmission du document <l>, bénéficiaire de sa rédaction <b>) et interne (notamment linguistique, cf. *infra* 6). Une fois identifié le rédacteur, il est possible d'indiquer à quel groupe social il appartient et dans quel cadre géographique il doit être placé ; mais pour cette dernière donnée, une réflexion sur la segmentation de l'espace géolinguistique s'impose (est-ce que "Salival" est une entité géolinguistiquement pertinente ? Nous avons choisi ici un espace plus large, la "Lorraine ducale", correspondant à la diffusion potentielle, supposée pour le document).

Actuellement, nous essayons d'appliquer des critères semblables au *Corpus d'Amsterdam*, corpus de 290 textes littéraires du XIIIe siècle : la valeur linguistique de cet ensemble important gagnera considérablement par l'ancrage plus précis des différents témoins dans le diasystème de l'ancienne langue (cf. Stein/Gleßgen 2005).

5. Analyse linguistique de base

Pour l'analyse linguistique du corpus textuel constitué, nous avons développé grâce à une collaboration transdisciplinaire un programme pilote, *Phoenix*[6]. Le premier élément de *Phoenix* est un lemmatisateur-xml qui permet de regrouper les formes lexicales et onomastiques des textes sous des lemmes ou dans des ensembles graphématiques ou morphologiques. Ce lemmatiseur fonctionne, dans un premier temps, sans outils complémentaires (comme des lexiques de référence). Il repose sur la segmentation des mots graphiques, mais l'interprétation des phraséologismes demande encore une réflexion complémentaire. La procédure est classique : le lemmatiseur réunit les formes dans un index alphabétique com-

[6] Les programmes ont été écrits dans Tustep par Matthias Kopp (université de Tübingen) et, dans un deuxième temps, par Matthias Osthof (université de Zurich) ; ils ont été conçus par nous trois lors de mises en commun de réflexions aussi prenantes que vives.

plet qui donne les mots dans la ligne de contexte (Index-KWIC [*Key Word in Context*]). Le travail linguistique consiste à désambiguiser les formes et à leur attribuer, par exemple, un lemme lexical.

Le lemmatiseur contient néanmoins quelques fonctionnalités pour faciliter le travail : il distingue automatiquement les noms propres (écrits avec majuscules) des noms communs (écrits avec minuscules, sauf en début de phrase) ; et il permet la définition d'équivalences graphématiques qui réduisent notablement la variance des formes (p.ex. des équivalences fonctionnelles : *abé*, *abbé*, *abei*, *abbei*, *abbey* reflètent le même lemme ; des consonnes doubles comme *cc/c*, *dd/d*, *ff/f*, *bb/b*, des groupes consonantiques latinisants comme *cq/q* ou *ct/t*, des homophones comme *en/an*, *y/i* et *-z/-s* ou des variations grapho-phonétiques régionales comme *-ei/-é*, *np/mp* et *w/g*; cf. Gleßgen 2003).

Par ailleurs, le lemmatiseur permet à tout moment de voir le contexte long des formes jusqu'au contexte complet d'une charte (cf. Gleßgen/Kopp 2005). Les interrogations peuvent porter sur l'ensemble du lexique (ou des noms propres) mais aussi sur un choix de mots selon la fréquence (haute ou faible) ou sur des suites de lettres (p.ex. réunir toutes les formes qui contiennent *able*, *aible*, *auble*, *aule* et *auvle* ou le suffixe *-ment*).

Les formes identifiées peuvent ensuite être retenues pour un balisage qui attribuera alors à un ensemble donné un lemme ("abbé") ou une qualité graphématique ("able3" = 3e type graphématique qui correspond à la forme *auble*). Dans le texte de base, toutes les formes en question reçoivent une balise neutre qui contient un numéro d'identification (p.ex. *Conue chose soit a-toz que li <wn n =„328“> abes </wn> et li chapitles de Salinvas*; "wn" = *Wortnummer*). Parallèlement un fichier-index répertorie les qualités attribuées à la forme en question (p.ex. *<wn> 328 </wn> <src> abes </src> <lex f=„c“> abbé </lex>*; "src" = *forme source*, "lex" = lemme, "c" = nom commun [à la différence des noms de lieux ou de personne]). Les données du fichier-index peuvent être projetées à tout moment sur le fichier textuel.

Ce lemmatiseur contient essentiellement deux qualités qui le distinguent de la plupart des autres lemmatiseurs actuellement en usage : d'abord, il reproduit de manière très fidèle le travail traditionnel (ou manuel) du philologue qui prépare l'analyse linguistique d'un texte. Ensuite, il garantit à tout moment un lien vivant entre les données interprétatives et le texte de base : celui-ci peut, par conséquent, être corrigé à tout moment, sans grande complication. Puisque l'analyse linguistique mène fatalement à des corrections du texte, la possibilité d'intervention après la lemmatisation est très précieuse.

Par ailleurs, si le lemmatiseur fonctionne en principe sans lexique de référence, il peut parfaitement intégrer, dans un deuxième temps, de tels outils ou un balisage morphologique obtenu par des moyens automatiques (p. ex. à l'aide du TreeTagger, cf. Stein/Gleßgen 2005).

Un deuxième programme permet ensuite de structurer plus en détail et de commenter les données balisées. Les formes réunies sous un lemme peuvent être organisées d'après leur caractéristiques morphologiques ("adj.sg.fém.", "v.tr.prés.3" etc.) et définies ; ils peuvent recevoir une description détaillée, en termes de lexicographie historique, d'étymologie et d'histoire linguistique. Les éléments diasystématiques sont à tout moment présents pour les différentes formes et peuvent intervenir dans la structuration. L'application de ce deuxième programme correspond dans les faits à la création de bases de données lexicologiques, ono-

mastiques, graphématiques et morphologiques qui s'enrichissent de chaque corpus traité. Cette partie du logiciel *Phoenix* vient à peine d'être achevée et sera utilisée à partir des prochains mois pour la mise en place de glossaires et analyses graphématiques.

Un troisième programme en cours d'élaboration gèrera enfin l'exportation des données lexicologiques et onomastique sous forme de glossaires ou de dictionnaires.

6. Analyses ultérieures

Une fois les trois programmes de *Phoenix* bien constitués, ils pourront être utilisés pleinement pour des réalisations et analyses linguistiques. Pour la série des *Plus anciens documents*, la lemmatisation et la description lexicographique a commencé avec les corpus de la Meurthe-et-Moselle et de la Moselle (cf. Matthey, ici) ; chaque nouveau corpus utilisera toutes les définitions et commentaires déjà introduits dans la base de données lexicologique ; toute nouvelle forme ou tout nouveau sens élargira alors cette base qui finira dans un temps très raisonnable par représenter un supplément conséquent au dictionnaire de Godefroy.

Le traitement des données onomastiques des différents volumes de la série se poursuivra parallèlement, à partir de 2006. Étant donné les lacunes considérables dans ce domaine, la base développée sera d'une utilité particulière pour l'histoire linguistique. Les trois bases, lexicologique, toponymique et anthroponymique, seront naturellement liées par des renvois réciproques pour les formes délexicales et déonymiques.

La base lexicologique permettra aussi l'intégration d'autres matériaux, notamment ceux du *Corpus d'Amsterdam*, déjà mentionné. La lexicographie existante de l'ancienne langue française (FEW, TL, Gdf, DEAF, AND) recevra ainsi une nouvelle dimension textuelle et contextuelle.

Le volet graphématique et morphologique de *Phoenix* servira dans un premier temps à identifier précisément les rédacteurs des différentes chartes. Une étude pilote des lieux d'écriture dans le corpus de Meurthe-et-Moselle (Gleßgen [s.p.]) montre qu'il existe des différences notables entre les choix graphématiques des comtes de Bar, des ducs de Lorraine, des scriptoria ecclésiastiques et des scribes libres travaillant pour la petite noblesse et les villages : la chancellerie des comtes de Bar fait appel à des formes de haut prestige, en correspondance avec le français central (*estable*, *le*, *lettres*) ou avec le picard (*avera*) ; elle se caractérise par une grande homogénéité des graphies (*ceus*) et des choix “modernes” (*prendre*). Les scribes libres au contraire choisiront facilement des graphies régionales lorraines (*estaule*, *lo*, *lattres*, *avra*, *ceas*, *penre*). La chancellerie des ducs de Lorraine occupe une position intermédiaire (variance entre *avra* et *avera*, *estable* et *estauble* ; homogénéisation partielle, sans formes régionales marquées), de même que les scriptoria épiscopaux et des grandes abbayes (formes d'un prestige linguistique très variable : *estable*, *estauble*, *estaule*, *estauvle*, etc.).

L'étude systématique des rédacteurs dans les différents corpus permettra l'établissement d'un réseau des lieux d'écriture dans la *Langue d'oïl* qui fournira un nouvel ancrage pour l'étude des textes et des formes dans le diasystème de l'ancienne langue.

Les finalités premières de *Phoenix* s'arrêtent là. Le programme a été conçu pour permettre l'édition et l'analyse linguistique de textes anciens sur la base d'un encodage rigoureux de type xml. D'autres outils permettront des interrogations d'histoire linguistique quantificative dans tous les domaines : la mise en place d'une normalisation graphématique (cf. Matthey, ici), les distributions statistiques du lexique, la structure textuelle ou la conception de la phrase médiévale. Pour ces fins, nous avons l'intention d'utiliser des outils de référence comme X-Query utilisé avec grand profit dans le AND (cf. Beddow, ici).

Tous les pas dans cette philologie informatique coûtent un investissement en termes de temps et de réflexion notablement plus élevé que dans la philologie traditionnelle qui s'appuyait sur l'intuition et la flexibilité d'un savoir disciplinaire solide. Mais la philologie informatique ouvre néanmoins de nouvelles perspectives qui méritent ce nouvel engagement.

7. Bibliographie

Beddow, Michael (2005) : *L'*Anglo-Norman On-line Hub *: une présentation technique*. Dans ce volume, # - # .

Gleßgen, Martin-D. (2003) : *L'élaboration philologique et l'étude lexicologique des Plus anciens documents linguistiques de la France à l'aide de l'informatique*. In : Duval, Frédéric (éd.) : *Frédéric Godefroy. Actes du X^e colloque international sur le moyen français*, Paris, Ecole des Chartes, 371-386.

- (1995) : *Diskurstraditionen zwischen pragmatischen Regeln und sprachlichen Varietäten*. In : Schrott, Angela / Völker, Harald (éd.), *Historische Pragmatik und historische Varietätenlinguistik in den romanischen Sprachen*, Göttingen, Universitätsverlag, # - #.

- (sous presse) : *Les 'lieux d'écriture' et leur identification dans les documents lorrains du XIII^e siècle*. In : *RLiR*.

- / Kopp, Matthias (2005) : *Linguistic annotation of texts in non-standardized languages : the program procedures of the tool PHOENIX*. In : Pusch / Kabatek / Raible 2005, 147-154.

Matthey, Anne-Christelle (2005) : Dans ce volume, # - # .

Pusch, Claus / Kabatek, Johannes / Raible, Wolfgang (éd.) (2005), *Romanistische Korpuslinguistik II : Korpora und diachrone Sprachwissenschaft / Romance Corpus Linguistics II : Corpora and Diachronic Linguistics*, Tübingen, Narr.

Stein, Achim / Gleßgen, Martin-D. (2005) : *Resources and Tools for Analyzing Old French Texts*. In : Pusch / Kabatek / Raible 2005, 135-145.