Martin-Dietrich Gleßgen

Élaborations philologiques et linguistiques sur la base de corpus textuels en français ancien – architecture du projet

1. Les projets de recherche

La présentation suivante essaie de mettre au clair l'architecture d'un ensemble de projets médiévistes qui nous occupent depuis 1997 et qui ont comme but général de trouver une synthèse entre la philologie, l'historiographie linguistique et l'informatique. L'historiographie linguistique et la philologie connaissent, depuis la fin du XX^{e} siècle, une transformation radicale à cause des développements informatiques qui renouvellent leurs outils disciplinaires. Puisque cette transformation médiale affecte l'intégralité des sciences et du monde de la communication, chacune des sciences doit forger sa propre synthèse entre les nouveaux outils et ses traditions méthodologiques. Si la linguistique actuelle est riche en innovations, d'ordre interprétatif et d'ordre pratique, la révolution informatique a néanmoins un impact particulier, notamment dans la recherche historique et philologique; celle-ci a comme objets immédiats des textes écrits linéaires, s'inscrivant donc parfaitement dans la logique d'interprétation des ordinateurs, elle aussi linéaire.

L'intégration des deux mondes se heurte toutefois à de notables difficultés pratiques et théoriques, à commencer par les compétences très diverses auxquelles ils font appel. Il y a une rupture presque totale entre les personnes qui ont des connaissances et intérêts philologiques et historiques –comme l'auteur de ces lignes– et celles qui ont de véritables capacités de programmation. Par ailleurs, il s'agit de joindre deux univers dont chacun se trouve aux prises avec des processus accélérés de transformation. De très nombreux chercheurs travaillent, parallèlement, à l'interface entre la linguistique (historique) et l'informatique, ce qui brouille le regard. La réflexion scientifique est enfin perturbée par l'absence d'une formation de doctrine, de tentatives de hiérarchisation et d'évaluation dans les outils et les méthodes.

Nos projets de recherche philologique se sont construits dans ce monde en pleine évolution; par conséquent, ils évoluent, eux aussi, au fur et à mesure qu'ils se construisent.

À l'heure d'aujourd'hui, il s'agit de trois projets d'une ampleur très différente:

(1) L'édition et l'analyse linguistique des *Plus anciens documents linguistiques de la France* constituent le noyau dur des différentes réalisations (AncDocFr). Après le décès de Jacques Monfrin, Françoise Vielliard, Olivier Guyotjeannin et l'auteur de ces lignes se sont associés pour répertorier les documents français du XIII^{e} siècle conservés dans les différents départements de la langue d'oïl, pour les éditer nouvellement ou, si possible, pour reprendre d'anciennes éditions. C'est un projet aux vastes perspectives, étant donné le

faible nombre de recherches linguistiques menées dans le passé sur des textes documentaires. S'ajoutent les qualités particulières des chartes d'être souvent transmises sous forme originale, d'être facilement datables et de connaître une nette variation géolinguistique et aussi diastratique.

(2) Un deuxième projet concerne la meilleure exploitation du *Corpus d'Amsterdam* de textes littéraires du XIII[e] siècle, établi sous la direction d'Antonij Dees et de Pieter Van Reenen. Avec Achim Stein et Pierre Kunstmann, nous avons repris ce corpus important et l'auteur de ces lignes a pris la part de la critique et de l'évaluation philologique des textes, dont deux tiers (62%) ont été transcrits à partir d'un manuscrit défini. Le projet du *Nouveau Corpus d'Amsterdam* (ou: *Sources et outils pour le français ancien*, SOFA) est moins prenant que les AncDocFr mais ouvre de très belles perspectives, notamment dans la comparaison des textes littéraires et documentaires de la même époque.

(3) Ces deux projets interviennent ensemble dans le cadre du *Consortium international pour les Corpus du Français Médiéval* (CCFM). Le consortium a pour objectif de réunir différents corpus sur une même plate-forme pour permettre, à terme, des interrogations communes. Dans un premier temps, il s'agit, en dehors des *Plus anciens documents* et du *Nouveau Corpus d'Amsterdam*, des textes réunis par le *Laboratoire de français ancien* d'Ottawa, de la *Bibliothèque du Français médiéval* (ENS, Lyon) et la base textuelle du *Dictionnaire du Moyen Français* (ATILF, Nancy).

L'idée directrice et fédératrice entre ces différents projets est d'ordre philologique: la linguistique de corpus appliquée à des textes anciens doit se valoir d'un soin particulier de la base philologique. Un ancrage diasystématique solide des formes linguistiques nécessite notamment des transcriptions (semi-)diplomatiques de manuscrits définis.

Avec le temps, l'architecture de ces projets, centrés surtout autour des *Plus anciens documents*, a développé un degré de complexité notable au point de devenir opaque pour tout observateur extérieur. Il nous a semblé utile, par conséquent, de décrire leur organigramme actuel en présentant (2) l'organisation administrative et les personnes impliquées dans les projets, (3) la structure des données textuelles et linguistiques ainsi que (4) les objectifs scientifiques et les méthodes auxquelles nous faisons appel.

2. La structure externe et les personnes impliquées

Les AncDocFr sont rattachés à l'université de Zurich et à l'École Nationale des Chartes. L'apport financier de loin le plus puissant provient actuellement du Fonds National Suisse dans le cadre du Pôle de Recherche National (PRN) «*Médialité. Transformation, changement et savoir médiaux*» (Projet D 4: *Écrit documentaire et élaboration linguistique du français au Moyen Âge tardif* [intitulés allemands: Nationaler Forschungsschwerpunkt, NFS: *Medialität. Medienwandel, Medienwechsel, Medienwissen*. Projet: *Geschäftsschrifttum und Sprachausbau im französischen Spätmittelalter*]).

Ce cadre interdisciplinaire met en avant l'impact de la médialité sur l'évolution linguistique. Les chartes se trouvent à l'interface entre l'écrit sous forme d'un support bien

particulier et leur utilisation dans un contexte pragmatique défini (= situation médialisée) et avec des finalités et effets politiques divers. Le rattachement du projet au pôle de médialité a eu comme effet un débat suivi, parfois surprenant et toujours fructueux avec des collègues et thésards historiens, juristes et littéraires, souvent germanistes travaillant sur des sources textuelles, iconographiques et architecturales sous diverses optiques.

Dans ce cadre se placent trois thèses actuellement en cours sur les chartes françaises, celle de Julia Alletsgruber, de lexicologie (cf. *infra* 4.1, 5.2.2), et de Claire Muller, de syntaxe (cf. *infra* ib.), ainsi que celle de Dumitru Chihaï, ciblée sur les lieux d'écriture. Le FNS finance par ailleurs le travail d'un programmateur à temps partiel, Matthias Osthof, qui est responsable, avec Matthias Kopp (Tübingen), de la part informatique de notre logiciel philologique *Phoenix* (cf. *infra* 4.3). Par ailleurs, notre secrétaire au *Romanisches Seminar* de Zurich, Cristina Solé, intervient dans la saisie et l'adaptation de transcriptions anciennes (notamment les chartes du Nord, cf. *infra* 3.1).

Le *Nouveau Corpus d'Amsterdam* a fait l'objet d'un financement binational, obtenu par les universités de Stuttgart (Achim Stein) et Ottawa (Pierre Kunstmann); à Zurich, nos assistants Xavier Gouvert et Claire Vachon ont investi un temps notable dans l'évaluation philologique du corpus.

Les *Corpus du français médiéval* n'ont pas, jusqu'ici, d'existence structurelle propre. Les partenaires se rencontrent une fois par an depuis la constitution du consortium, à tour de rôle à Ottawa, Stuttgart / Lauterbad, Nancy, Lyon et Zurich.

En dehors des partenaires déjà mentionnés, la contribution la plus importante revient à Paul Videsott, de l'université de Bolzano / Bressanone. Il a établi, pendant l'année 2005 / 2006, un inventaire de 3.600 chartes françaises et latines émanant de la royauté française et de l'entourage royal immédiat (cf. *infra* 5.2.5). D'autres collaborations sont en vue puisque c'est un domaine de la recherche qui permet des segmentations cohérentes et opératoires et laisse la place au concours des forces de nombreux chercheurs.

Autant les financements que les personnes impliquées dans ces projets sont soumis à des fluctuations continuelles, en fonction de facteurs les plus divers. Nous avons commencé l'étude des *Plus anciens documents* à Strasbourg, en construisant autour de ce projet une petite équipe de travail dans le cadre de la *Maison interuniversitaire des Sciences humaines en Alsace* (MISHA), équipe qui réunissait nos confrères André Thibault et Dominique Gerner, le médiéviste Benoît Tock et le littéraire italianisant Johannes Bartuschat. Notamment, l'aide du diplomatiste qu'est Benoît Tock nous a été alors d'un secours inestimable.

Le cadre zurichois est bien plus vaste et il a apporté de nouveaux interlocuteurs comme le germaniste littéraire Christian Kiening, directeur du Pôle de Recherche, les médiévistes Martina Stercken et Simon Teuscher, l'historien du droit Andreas Thier ou la collègue germaniste Elvira Glaser. Le projet du FNS est financé de 2005 jusqu'en 2009; il peut y avoir une suite mais ce n'est jamais établi, ce qui complique considérablement le planning.

Il faut souligner par ailleurs la participation ininterrompue de jeunes chercheurs depuis 1999 (dans l'ordre chronologique: Jason D. Stein, Frédérique Gisquet, Delphine Harmand et Séverine Constans avec des mémoires de maîtrise et Anne-Christelle Matthey avec sa thèse de doctorat).

Un facteur qui pourrait s'avérer comme élément stabilisateur est l'enseignement. Le nouveau cursus de *BA* prévoit pour les études de français à Zurich des cours en paléographie, en lexicographie historique et en philologie informatique. Cela permet de réorganiser les choix méthodologiques opérés dans les projets de recherche pour l'enseignement, ce qui contribuera sans doute à une meilleure assise de la doctrine scientifique.

3. La structure des données textuelles

L'élaboration des données textuelles et linguistiques se caractérise par une clarté de départ infiniment plus grande que l'organisation externe, de même que par une évolution simple et linéaire. La science est plus rigoureuse et plus construite que les aléas institutionnels.

3.1 Les données textuelles des *Plus anciens documents*

L'idée de base de Paul Meyer, poursuivie par Clovis Brunel, puis par Jacques Monfrin, a été celle d'étudier, département par département, les collections des anciens documents en langue vernaculaire. Par cette voie, à terme, l'intégralité des documents intéressants auront été vus, surtout si l'on prend aussi en considération les documents conservés en Belgique, en Suisse romande et en Angleterre. Il sera possible alors de regrouper nouvellement les documents par leur lieu de genèse médiéval en vue de leur exploitation linguistique.[1]

Pour tous les documents, il s'agit de les transcrire d'après des critères définis et sous un format XML cohérent (cf. *infra* 5.2.1, 2003a; 2007a), d'identifier les paramètres diasystématiques et de les photographier pour permettre une édition web avec l'image à l'appui.

Le tableau suivant répertorie les quatorze ensembles de la langue d'oïl en chantier, de même que cinq ensembles en suspens pour lesquels existe des travaux préliminaires. Il ne prend pas en considération les départements du Midi −dont certains sont pourtant à l'étude− ni la Suisse romande, le Luxembourg ou la Belgique qui sont ou ont été traités par d'autres collègues. Il indique par ailleurs les éditeurs et adaptateurs principaux et le nombre de documents en cours d'édition.

Département	Éditeur(s) [adaptateur(s)]	Nombre de documents
I. Volumes publiés		
Aube, SM, Yonne	Coq	103
Haute-Marne	Gigot [Tock, Chihaï]	142
Oise	Carolus-Barre [Tock; Grübl]	202
Vosges	Lanher [Trotter]	285

[1] Il s'avère toutefois que de nombreuses chartes sont conservées encore aujourd'hui dans leur région d'origine, ce qui facilite les analyses déjà dans les premières phases du travail.

II. Volumes nouveaux en cours		
Jura	Muller	105
Marne	Chihaï	230
Meuse	Matthey	250
Moselle	Pitz	180
Nièvre	Alletsgruber	30
Haute-Saône	Muller	135
Saône-et-Loire	Alletsgruber	95
Chancellerie royale	Videsott	150 [+ 350]
III. Volumes ms. en réélaboration		
Meurthe-et-Moselle	Arnod; Gleßgen	290
Nord	Mestayer [Solé]	350
IV. Volumes ms. en suspens		
Côte-d'Or	Neveu	270
Aisne	Grégoire-Ollivier [Harmand]	220
Doubs	Lefèvre; Beaugendre [Camps]	250
Pas-de-Calais	Bougard	130
Somme	Estienne	100

Nous pouvons raisonnablement espérer mettre en ligne les dix départements en chantier à partir de 2008, à commencer par les volumes de la Meurthe-et-Moselle, de la Meuse et, sans doute, de la Haute-Marne qui comptabilisent presque 700 chartes. Une fois ce premier ensemble réuni, il comportera environ 2500 chartes. Notons que le degré d'élaboration des volumes déjà publiés sous forme papier sera moins développé que celui des nouveaux volumes puisque les principes éditoriaux et de description ont considérablement évolué depuis les années 1980.

Le tableau fait ressortir certaines difficultés qui résident dans l'harmonisation et l'élaboration cohérente de ces collections. De nombreux auteurs et adaptateurs, aux statuts les plus divers, interviennent sur des éditions qui se trouvent, elles, dans un état d'élaboration et d'avancement très variable. En dernière instance, l'élaboration d'une charte à partir d'une transcription dactylographiée des années 1980 d'après nos critères prend autant de temps que la nouvelle édition d'une charte non publiée qui s'inscrit immédiatement dans la logique actuelle du projet. Par conséquent, l'avenir des volumes manuscrits en suspens n'est pas plus établi que celui des départements de l'Ouest de la France ainsi que ceux du Midi.

Pour ne pas surcharger la lecture, le tableau ne comporte pas les dates extrêmes, pourtant significatives puisque les documents ne couvrent pas partout la même époque. Pour la Meurthe-et-Moselle, très riche en documents français anciens, Michel Arnod avait arrêté les transcriptions, à juste titre, en 1265, comptabilisant déjà jusque là 290 chartes. Pour Douai, le lieu d'écriture le plus précoce, Monique Mestayer a arrêté la transcription de 500 chartes à 1270 et nous avons décidé de saisir seulement les 350 premières, étant donné le temps de travail nécessaire pour l'intégration des nouveaux critères élargis d'édition. En revanche, pour les départements franc-comtois et bourguignons, les dates extrêmes ont dû être repoussées jusqu'en 1290 voire 1330 pour pouvoir documenter les débuts de l'écrit français dans la région. Les collègues travaillant sur les documents de la Suisse Romande sont allés jusqu'au milieu du XIV^{e} siècle. Ces décalages créeront des difficultés

d'interprétation mais ils correspondent à une réalité historique multiforme. Nous pourrons étudier dans tous les cas les débuts de l'écrit documentaire dans les régions en question et identifier les voies de diffusion des modèles linguistiques.

3.2 L'élaboration informatique des *Plus anciens documents*

L'édition des *Plus anciens documents* est accompagnée, dès les débuts, d'analyses lexicologiques, onomastiques et graphématiques. L'étude des lexèmes et des toponymes est indispensable ne serait-ce que pour la bonne compréhension et l'établissement du texte. Les éléments graphématiques et, dans une certaine mesure, morphologiques contribuent avec l'analyse paléographique à l'identification du lieu de genèse des documents (= ‹lieux d'écriture› ou ‹rédacteurs›). Ce sont donc des analyses ‹primaires›, incontournables dans une entreprise du genre.

Les données lexicologiques, onomastiques, graphématiques et morphologiques sont répertoriées dans une base de données à part qui est interdépendante avec la base de données textuelles. La configuration informatique des différentes bases contient précisément trois ensembles, (1) les données textuelles, (2) les fichiers-index et (3) la base de données interprétatives.

Les données textuelles sont organisées par fichiers individuels regroupant des ensembles thématiques (par ex. un fichier pour les 289 chartes de la Meurthe-et-Moselle, un autre pour les 250 chartes de la Meuse etc.). L'édition de chaque charte est accompagnée par un tableau analytique individuel, regroupant toutes les informations pertinentes pour son ancrage diasystématique et historique (cf. *infra* 5.2.1, 2003a; 2007a).

(1) fichier textuel:
<gl>
<an>[date, rédacteur, etc.]</an>
<txt> Le chapitre et li abbes de Salival (...) </txt>
</gl>

Les formes du fichier textuel sont enrichies, au fur et à mesure, par des informations linguistiques diverses: le lemme (lexical ou onomastique) auquel une forme peut être rattachée, sa catégorisation et description morphologique, les caractéristiques graphématiques ou, éventuellement, des variables dans le marquage morphologique (forme de l'article etc.). Ces informations sont placées, pour des raisons de gestion informatique, dans le fichier-index; le lien entre la forme et les informations linguistiques est établi par une balise portant un numéro univoque (wn).

(1) fichier textuel:
<gl>
<an>[date, rédacteur, etc.]</an>
<txt> Le chapitre et li <wn n="**1**">**abes**</wn> de Salival (...) </txt>
</gl>

(2) fichier-index:
<wn>**1**</wn> <src>**abes**</src> <lex f="c"abbé</lex> <graph f="(...) <morph f="(...) <sem f="(...)

Il est facile, grâce à un programme spécifique, d'importer les informations linguistiques dans le fichier textuel pour permettre des interrogations en texte plein.

(1+2) fichier textuel avec les informations du fichier-index:
<gl>
<an>[date, rédacteur, etc.]</an>
<txt> Le chapitre et li <wn n="1"> ***<idx>*** *<lex f="c"abbé</lex> <graph f="(...)* *<morph f="(...)* *<sem f="(...)* ***</idx>*** abes</wn> de Salival (...) </txt>
</gl>

Enfin, dans la base interprétative de données, les éléments regroupés auparavant sont réunis (par ex. toutes les formes appartenant à un lemme ou toutes les formes pour une variable graphématique) et peuvent être classés, décrits et commentés. La structure de cette base de données n'est pas banale mais elle est très clairement dessinée. Notamment, elle permet pleinement l'élargissement de la base textuelle de données, de même que des corrections dans le texte au cours de l'analyse (cf. Völker en ce volume).

3.3 Le Nouveau Corpus d'Amsterdam

Notre intervention dans le *Nouveau Corpus d'Amsterdam* est beaucoup plus modeste. Dans une première phase, nous avons essayé d'identifier précisément les 301 textes et de dater et localiser autant les manuscrits en question que les oeuvres à leur base. Ce travail, effectué surtout par Xavier Gouvert –dont les réalisations demandaient toutefois une attention constante– et, de manière parfaitement fiable, par Claire Vachon, s'est avéré bien plus épineux mais aussi bien plus utile que cela ne pourrait sembler. Il est possible maintenant, grâce à la programmation d'Achim Stein, d'interroger les formes du corpus d'après ces quatre paramètres (dates / lieux des oeuvres ou des manuscrits), ce qui donne une nouvelle sécurité de jugement à cet excellent outil linguistique. Nous avons aussi commencé une critique de la nature et de la qualité des éditions individuelles, mais cette opération devra se poursuivre dans une deuxième phase et prendre en considération les photographies des manuscrits.

Jusqu'ici, la localisation des manuscrits suit simplement les indications de la tradition philologique, condensée pour la plupart dans le *Complément bibliographique* du DEAF. Dans une deuxième phase, nous envisageons d'étudier donc plus en détail le phénomène de l'espace en nous aidant des résultats scriptologiques obtenus à partir des chartes.

Par ailleurs, nous projetons, avec Pierre Kunstmann et Achim Stein, d'intervenir par la suite dans la constitution du corpus, en enlevant des textes mal édités et en ajoutant de nouvelles parties de textes déjà présents ainsi que de nouveaux textes. En cela, le projet suppose une réflexion plus globale sur la transmission manuscrite du XIII^{e} siècle dans le domaine des textes non-documentaires.

Les données textuelles du *Nouveau Corpus d'Amsterdam* pourront être intégrées dans la structure informatique élaborée pour les *Plus anciens documents* puisqu'elle est clairement construite et extensible. Il pourra aussi l'enrichir grâce à son balisage des catégories

grammaticales et des lemmes. C'est un projet d'avenir qui s'inscrira en même temps dans l'optique des *Corpus du français médiéval*.

3.4 Les formes de publication prévues

La publication des bases textuelles des *Plus anciens documents* est prévue, essentiellement, sous une forme informatique, tout comme les textes du *Corpus d'Amsterdam*, déjà disponibles sur le web. Nous avons renoncé, pour la collection complète, à une publication-papier systématique qui se serait inscrite dans la collection des *Plus anciens documents linguistiques de la France*. Le fait de travailler à la fois sur plusieurs départements voisins dont les textes de rédaction se chevauchent aurait rendu incohérente la publication traditionnelle par volumes, dédiés chaque fois à un département. La publication-web permettra néanmoins une impression de qualité de toutes les chartes, organisées d'après des critères divers.

La publication-web sera évolutive. Chaque nouveau sous-ensemble départemental augmentera la base textuelle et, par là, les témoignages pour les différents lieux d'écriture. L'édition électronique permettra aux utilisateurs de choisir entre une présentation diplomatique ou interprétative des textes; la version qui sera proposée pour l'impression prévoit notamment un système mixte. La forme informatique intégrera les photographies des documents, ce qui en augmente l'utilité. Enfin, la version-web des textes sera en lien avec les bases de données linguistiques; celles-ci resteront modestes aux débuts mais évolueront au fur et à mesure, sous forme de répertoire voire de dictionnaire électronique.

4. Les données et interprétations linguistiques

4.1 Les réalisations linguistiques et philologiques ciblées

L'élaboration des données textuelles s'accompagne d'emblée d'éléments d'interprétation linguistiques. Leur mise en oeuvre capitalise l'investissement majeur à l'intérieur de nos projets, autant en termes de temps qu'en termes de réflexion. Les quatre bases de données interprétatives en construction (lexicologique, onomastique, graphématique, morphologique) sont réunies dans un seul fichier informatique mais correspondent, naturellement, à différentes interrogations et font appel à différentes méthodologies:

(1) La variation graphématique et morphologique est interprétée dans une optique scriptologique dont l'analyse est indispensable pour l'identification des lieux d'écriture. Les éléments d'analyse graphématique entreront dans différents travaux en chantier et déboucheront éventuellement sur une étude monographique d'interprétation scriptologique que nous projetons avec Paul Videsott.

(2) Le vocabulaire des documents est lemmatisé (mots lexicaux) et étudié partiellement de façon monographique. La première étude, de Julia Alletsgruber, concerne le vocabulaire du monde agricole, particulièrement présent dans les chartes; les fiches lexicologiques, très

développées, feront partie intégrante de sa thèse. Sur la base de certaines de ces entrées, nous avons étudié le marquage diasystématique dans le vocabulaire médiéval (cf. *infra* 5.2.4, 2006a; cf. aussi 2004 et 5.2.3, [sous presse]). Par ailleurs, nous avons prévu pour les collections en cours de l'Est de la France un volume réunissant toutes les entrées de glossaires lexicologiques. Ce macro-glossaire raisonné formera un supplément substantiel du *Dictionnaire* de Godefroy.

(3) Les noms de personne et, surtout, les toponymes sont traités d'après des méthodes lexicologiques, en vue d'un premier dictionnaire onomastique informatisé pour la Galloromania, petit mais évolutif. Parallèlement au volume dictionnairique, un deuxième volume imprimé réunira les entrées onomastiques contenues dans le même corpus textuel.

(4) La syntaxe est analysée dans un premier temps dans une optique de structuration textuelle en prenant en considération aussi la ponctuation et les majuscules dans les manuscrits. C'est le sujet de la thèse en cours de Claire Muller. Si nous devions construire par la suite une base syntaxique de données, ce serait probablement un projet commun avec Achim Stein.

Les analyses linguistiques sont accompagnées d'une tentative de description paléographique des supports des chartes, dans l'optique de leur médialité. Notamment la thèse de Dumitru Chihaï proposera des éléments de structure pour l'analyse paléographique, indispensable pour l'identification des divers *scriptoria* et chancelleries.

4.2 Les objectifs méthodologiques en linguistique historique

Les éditions et analyses présentées auparavant poursuivent différents objectifs d'ordre méthodologique qui font et feront objet de publications thématiques. Voici les objectifs les plus saillants:

4.2.1 Contribuer à un standard satisfaisant dans les principes d'édition

La tradition romaniste nous livre les meilleurs modèles pour les éditions de texte mais, malheureusement, nous sommes loin de voir ces modèles appliqués dans chacune des éditions nouvelles qui voient le jour. Un élément novateur de nos projets est le principe de l'encodage double, intégrant dans une même édition (papier ou web) des éléments de structure (majuscules, ponctuation) médiévaux et modernes. Le jeu de balises que nous avons mis au point pour la saisie d'un texte pourra être utile à d'autres éditeurs qui voudront s'inspirer de ces mêmes normes.

Jusqu'ici, nous n'avons travaillé que sur des transcriptions d'un manuscrit défini. La prise en considération conjointe de plusieurs témoins textuels d'une même oeuvre complique considérablement les choses, notamment pour créer des liens opératoires entre les différentes versions. Dans ce domaine, la réflexion reste ouverte.

La variation des genres textuels introduit d'autres impératifs qui ne sont pas faciles à gérer. Nous considérons que la TEI est trop lourde pour pouvoir être utilisée par le gros des philologues élaborant des éditions de texte. Il faudra trouver des compromis viables, ce qui sera néanmoins nettement plus facile que la gestion des différentes versions d'un texte.

4.2.2 Préciser et rendre opératoire la notion de ‹lieux d'écriture› dans l'historiographie linguistique

L'analyse des chartes lorraines nous a amené, depuis 2002, à prendre en considération les ‹lieux d'écriture›, les *scriptoria* et les chancelleries, comme entité à part entière dans le processus de l'élaboration de l'écrit, au moins médiéval (cf. *infra* 5.2.3, 2008a). Sur la base de nos résultats, nous supposons que la dimension du lieu d'écriture doit être placée entre l'individualité des scribes et la norme abstraite d'une variété régionale ou sociale de la langue. Dans une certaine mesure, les scribes s'adaptent à des normes reconnaissables quand ils travaillent en un lieu d'écriture défini. Cela concerne autant la mise en page que le choix des graphèmes. Le comportement des scribes semble alterner, par ailleurs, selon le genre textuel. Mais sans déjà entrer dans la description de ces micro-normes, il nous semble important d'identifier le lieu et l'institution concrets auxquels se rattache la rédaction d'un texte.

Le lieu d'écriture contient en même temps une dimension dans l'espace et dans l'univers sociologique; le prestige linguistique d'une charte épiscopale de Toul est radicalement différent de celui d'une charte rédigée par un scribe libre travaillant dans la même ville. Cela complique d'ailleurs considérablement la représentation cartographique des formes de scripta médiévales, représentation qui doit intégrer une dimension diastratique.

La notion de lieux d'écriture sera aussi le noyau de la deuxième phase de travaux sur le *Nouveau Corpus d'Amsterdam*, avec la réflexion sur les genres textuels. Malheureusement, cette interrogation suppose la prise en considération des photos de manuscrits, ce qui demande un effort matériel conséquent.

4.2.3 Structuration linguistique des genres textuels

Les chartes constituent un ensemble relativement homogène parmi les genres textuels. Il est naturellement possible et utile de distinguer des sous-genres, mais la variation linguistique reste relativement faible. Le *Nouveau Corpus d'Amsterdam* introduit plus clairement la nécessité de réfléchir, pour les genres textuels, à une structuration linguistique possible qui dépasse les catégories traditionnelles, d'origine littéraire. Les genres n'existent pas de manière absolue mais dépendent de l'optique de l'observateur; d'un point de vue linguistique, il est légitime d'identifier comme genre un ensemble textuel qui se caractérise par un nombre défini de variables linguistiques. Cette recherche en est encore à ses origines mais elle est prometteuse.

4.2.4 Garantir un ancrage diasystématique détaillé des formes linguistiques

Nos éditions fournissent pour tout document une description diasystématique détaillée. Celle-ci comporte le temps (éventuellement dédoublé entre l'époque de la genèse d'une oeuvre et la rédaction d'un manuscrit), l'espace (lui-aussi dédoublé si nécessaire), le prestige social, le lieu d'écriture (qui synthétise dans un certain sens l'espace et le prestige

social), éventuellement le scribe ou l'auteur, de même que le genre textuel (qui traduit en même temps l'ancrage pragmatique et intervient dans le marquage diaphasique du texte).

Chaque forme linguistique individuelle est donc liée à ces paramètres qui permettent un ancrage diasystématique intégral (cf. *supra* 3.2). C'est seulement sur une telle base qu'il est possible de concevoir une historiographie linguistique qui prend en compte le diasystème de la langue ou, du moins, de l'écrit.

Notre jeu de balises prévoit une entrée pour chacun des paramètres diasystématiques et nos bases interprétatives de données les prennent pleinement en considération. L'introduction de ces paramètres dans le *Corpus d'Amsterdam*, jusqu'ici partielle, s'est avérée très utile pour les interrogations lexicologiques.

4.3 Les développements informatiques

Nous avons investi beaucoup d'énergie dans le développement du logiciel *Phoenix*, programmé concrètement avec Tustep par Matthias Kopp (Tübingen) et Matthias Osthof (Tübingen / Zürich). Ce logiciel réunit un programme d'édition, un lemmatiseur (qui permet aussi le regroupement de variables graphématiques et morphologiques), l'outil de saisie des bases de données interprétatives et un programme de représentation de ces bases.

Parallèlement, nous utilisons un éditeur XML pour la vérification du jeu de balises (par un schéma) et pour des interrogations à l'aide du langage X-Query.[2] Nous avons aussi réfléchi, avec Achim Stein, à des interfaces entre le TreeTagger, baliseur morphologique très performant, programmé par Heinrich Schmid et Achim Stein. Nous continuerons sans doute dans cette voie qui mènera vers des applications dans le cadre des *Corpus du français médiéval*.

Les outils développés sont maintenant disponibles et seront aussi téléchargeables à partir de 2009. Malgré leur utilité ils posent le problème d'être des réalisations exigeantes qui demandent un effort indéniable dans l'application. Un éditeur de texte et linguiste qui voudra utiliser ces outils devra:

- saisir ses données textuelles sous une forme XML ou les transformer en une forme XML,
- adapter cette forme XML à notre jeu de balises (cf. *infra* 5.2.1, 2003a; 2007a),
- installer Tustep sur son ordinateur et apprendre un certain nombre de commandes pour pouvoir lancer les programmes,
- comprendre le fonctionnement des deux programmes qui constituent *Phoenix*, le lemmatiseur et le programme des bases de données,
- éventuellement, se familiariser avec X-Query.

Pour faciliter l'accès à ces outils, nous sommes en train de rédiger des documents descriptifs. Dans cette même optique ont été réuni dans la bibliographie tous les textes rédigés jusqu'ici dans le cadre des projets (cf. *infra* 5.2); les publications sont complémentaires et permettent un accès à la logique des projets, y compris à celle des programmes informatiques.

---

[2] Malheureusement, la plupart des éditeurs xml sont de type commercial. Actuellement, nous travaillons avec le logiciel *Exchanger*, très satisfaisant et gratuit pour des usages universitaires.

Enfin, nous espérons que l'intégration de nos outils dans l'enseignement contribuera à une meilleure accessibilité. Toutefois, il est probable qu'il ne sera jamais possible de rester en deçà d'un certain degré de complexité dans l'utilisation d'outils informatiques dans l'historiographie linguistique.

4.4 Les finalités interprétatives ultérieures

Les finalités interprétatives se placent essentiellement dans l'optique du changement linguistique, entendu comme un outil pour comprendre le fonctionnement de la langue. Nos projets impliquent notamment le rôle de la variation diasystématique et celui des genres textuels dans le fonctionnement et dans le changement linguistiques. Ils permettront aussi de spécifier l'importance des paramètres médiaux pour les formes linguistiques et leur développement. Les chartes créent par ailleurs un lien particulièrement étroit avec l'infrastructure politique dans laquelle elles évoluent, ce qui permet de mieux cerner l'impact de l'ancrage pragmatique sur le développement des genres textuels et sur l'élaboration linguistique.

Il est vrai que le chemin à parcourir entre les recherches empiriques et leur interprétation finale est très long. Dans ce sens, le projet des *Plus anciens documents* a l'inconvénient de demander de notables efforts pour l'élaboration primaire des données linguistiques mais l'avantage de fournir des données suffisamment complexes pour permettre de véritables conclusions. Les lourdeurs inévitables de la matière garantissent en même temps des résultats novateurs et solides dans les différents domaines interprétatifs, au profit de l'historiographie linguistique tout court.

5. Références bibliographiques

5.1 Publications externes

Duval, Frédéric (ed.) (2003): *Frédéric Godefroy*. Actes du X^{e} colloque international sur le moyen français. Paris: Ecole des Chartes.

Gärtner, Kurt / Holtus, Günter / Rapp, Andrea / Völker, Harald (edd.) (2001): *Skripta, Schreiblandschaften und Standardisierungstendenzen* (Beiträge zum Kolloquium vom 16. bis 18. September 1998 in Trier). Trier: THF.

Gärtner, Kurt / Holtus, Günter (edd.) (2005): *Drittes Trierer Urkundensprachekolloquium* (20.-22. Juni 2001). Trier: THF.

Guyotjeannin, Olivier (ed.) (2006): *La langue des actes*. Actes du XI^{e} congrès de la Commission internationale de diplomatique (Troyes, 11-13 septembre 2003). Éditions en ligne de l'École des Chartes (elec: www.enc.sorbonne.fr/editions-en-ligne.html), n° 7.

Kunstmann, Pierre / Martineau, France / Forget, Danielle (edd.) (2003): *Ancien et moyen français sur le Web: enjeux méthodologiques et analyse du discours*. Ottawa: David.

Kunstmann, Pierre / Stein, Achim (edd.) (2007): *Le Nouveau Corpus d'Amsterdam*. Actes de l'atelier de Lauterbad, 23-26 février 2006. Stuttgart: Steiner.

Pusch, Claus / Kabatek, Johannes / Raible, Wolfgang (edd.) (2005): *Romanistische Korpuslinguistik II: Korpora und diachrone Sprachwissenschaft / Romance Corpus Linguistics II: Corpora and Diachronic Linguistics* (ScriptOralia, 130). Tübingen: Narr.

Pusch, Claus / Pfänder, Stefan / Raible, Wolfgang (edd.) (2008): *Romanistische Korpuslinguistik III: Korpora und Pragmatik / Romance Corpus Linguistics III: Corpora and pragmatics* (ScriptOralia). Tübingen: Narr.

Schrott, Angela / Völker, Harald (edd.) (2005): *Historische Pragmatik und historische Varietätenlinguistik in den romanischen Sprachen*. Göttingen: Universitätsverlag.

Völker, Harald. (en ce volume, section 13).

– / Schøsler, Lene / Gleßgen, Martin-Dietrich (edd.) (2007): *De la philologie aux nouveaux médias: éditions de textes – linguistique de corpus – analyse informatique du langage*. In: Trotter, David (ed.): Actes *du XXIV^e Congrès International de Linguistique et Philologie Romanes* (sept. 2004, Aberystwyth). Vol. 1: Section 2. Tübingen: Niemeyer, 285-480.

5.2 Publications dans le cadre du projet

[sauf indication contraire ou précision, l'auteur est M.-D. Gleßgen]

5.2.1 Présentations générales et questions d'édition

(2001): *Das altfranzösische Geschäftsschrifttum in Oberlothringen: Quellenlage und Deutungsansätze*. In: Gärtner / Holtus / Rapp / Völker (edd.), 257-294.
[= première présentation du projet, proposant des réflexions sur sa faisabilité].

(2003a): *L'élaboration philologique et l'étude lexicologique des Plus anciens documents linguistiques de la France à l'aide de l'informatique*. In: Duval (ed.), 371-386.
[= deuxième présentation en français, indiquant les critères d'édition et le jeu des balises].

(2005a): *Editorische, lexikologische und graphematische Erschließung altfranzösischer Urkundentexte mit Hilfe von TUSTEP. Stand der Arbeiten*. In: Gärtner / Holtus (edd.), 91-107.
[= id., avec peu de variations, en allemand].

(2007a): *Bases de données textuelles et lexicographie historique: l'exemple des Plus anciens documents linguistiques de la France*. In: Völker / Schøsler / Gleßgen / Di Girolamo (edd.), 373-380; = In: Aprile, Marcello (ed.): *Nuove riflessioni sulla lessicografia. Presente, futuro e dintorni del Lessico Etimologico Italiano*, Atti del Seminario di Lecce (21-22 aprile 2005). Galatina: Congedo, 157-167.
[= troisième présentation (publiée parallèlement en deux lieux), expliquant les critères d'édition et les éléments analytiques du tableau: les paramètres du diasystème et les genres textuels].

5.2.2 Lexicologie et syntaxe

(2003b): *La lemmatisation de textes d'ancien français: méthodes et recherches*. In: Kunstmann / Martineau / Forget (edd.), 55-75.
[= problèmes soulevés par la lemmatisation].

(2005b): *Linguistic annotation of texts in non-standardized languages: the program procedures of the tool PHOENIX*. In: Pusch / Kabatek / Raible (edd.), 147-154 (avec Matthias Kopp).
[= présentation du lemmatiseur].

Alletsgruber, Julia: en ce volume, section 13.
Muller, Claire: en ce volume, section 13.

5.2.3 Lieux d'écriture et régionalité

(2008a): *Les ‹lieux d'écriture› dans les chartes lorraines du XIII^e^ siècle*. In: *RLiR* 2008/2.
[= méthodologie de l'identification des ‹lieux d'écriture›].
(sous presse): *Église et puissance temporelle en terre lorraine: une charte épiscopale de Toul de 1237*. In: *Mélanges de linguistique et de philologie romanes*.
[= l'exemple des chartes épiscopales de Toul du XIII^e^ siècle].
Matthey, Anne-Christelle [s.a.]: *Les plus anciens documents linguistiques de la France: le cas du Département de la Meuse*. 1 vol. + 2 vol. d'édition. Univ. de Zurich (sept. 2006); cf. les présentations Pusch / Kabatek / Raible (edd.) et Völker, Harald / Schøsler, Lene / Gleßgen, Martin-D. (edd.).
[= régionalité et stéréotypie linguistiques (graphie, lexique, collocations)].

5.2.4 Genres textuels et diasystème

(2005c): *Diskurstraditionen zwischen pragmatischen Regeln und sprachlichen Varietäten*. In: Schrott / Völker (edd.), 207-228 [trad. esp. en préparation].
[= rôle des genres textuels dans le changement et le fonctionnement linguistiques].
(2006a): *Vergleichende oder einzelsprachliche historische Textwissenschaft*. In: Dahmen, Wolfgang *et al.* (edd.): *Was kann eine vergleichende romanische Sprachwissenschaft heute (noch) leisten?* (Romanistisches Kolloquium XX). Tübingen: Narr, 319-340.
[= identification des variables diasystématiques dans le lexique des chartes].
(2004): *Realia und Urkunden. Die Teilung eines lothringischen Stadthauses kurz nach 1400*. In: Gil, Alberto *et al.* (edd.): *Romanische Sprachwissenschaft. Zeugnisse für Vielfalt und Profil eines Faches*. Festschrift für Christian Schmitt zum 60. Geburtstag. Frankfurt a.M.: Lang, 423-447.
[= étude lexicologique/diasystématique et textuelle d'une charte lorraine de 1414].
(2008b): *Corpus historiques et pragmatique – genres textuels et variétés linguistiques*. In: Pusch, C. *et al.* (edd.): *Romanistische Korpuslinguistik III: Korpora und Pragmatik* / *Romance Corpus Linguistics III: Corpora and pragmatics* (ScriptOralia). Tübingen: Narr.
[= synthèse en français des articles précédents 2005 / 2006].

5.2.5 L'écrit documentaire dans l'histoire linguistique et dans l'histoire de la France

(2006b): *L'écrit documentaire dans l'histoire linguistique de la France*. In: Guyotjeannin (ed.): [18 p.] (http://elec.enc.sorbonne.fr/document328.html).
Videsott, Paul: en ce volume, section 13.

5.2.6 Le rôle de l'informatique dans la philologie et dans l'historiographie linguistique

(2006c): *Esigenze della tecnologia informatica nella filologia e lessicografia storica*. In: Schweickard, Wolfgang (ed.): *Nuovi media e lessicografia storica*. Atti del colloquio in occasione del settantesimo compleanno di Max Pfister. Tübingen: Niemeyer, 15-24.

(2007b): *Philologie und Sprachgeschichtsschreibung in der Romanistik: Die ‹informatische Wende›*. In: Stolz, Michael (ed.): *Edition und Sprachgeschichte*. Baseler Fachtagung 2.-4. März 2005. Tübingen: Niemeyer (Beihefte zur editio 26), 201-212.

5.2.7 Le Nouveau Corpus d'Amsterdam

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

(2007c): *La base textuelle du* Nouveau Corpus d'Amsterdam*: ancrage diasystématique et évaluation philologique*. In: Kunstmann / Stein (edd.), 51-84 (avec Xavier Gouvert).