Kunstmann, Pierre; Stein, Achim (2007): Le Nouveau Corpus d'Amsterdam. – Kunstmann, Pierre; Stein, Achim (ed.): Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006, Stuttgart: Steiner, 9-27.
(Preliminary, non final version)

LE NOUVEAU CORPUS D'AMSTERDAM

Pierre Kunstmann / Achim Stein

1 INTRODUCTION

1.1 Présentation du corpus

Le Nouveau Corpus d'Amsterdam est le résultat du remaniement du corpus de textes littéraires d'ancien français compilé par Anthonij Dees (Vrije Universiteit Amsterdam) et son équipe. Ce corpus a constitué la base de *l'Atlas des formes linguistiques des textes littéraires de l'ancien français* (Dees et al. 1987, cf. infra pour plus de détails). Les données informatisées qui restaient de ce projet ainsi qu'une partie de la documentation ont été sauvées par Pieter van Reenen, qui nous les a transmises en 1999 à l'Université de Cologne (invité par Peter Blumenthal, avec qui A. Stein préparait la version électronique du Tobler/Lommatzsch).[1] C'est donc grâce à l'intention de préserver ces données, à la collaboration de plusieurs chercheurs avec des intérêts et compétences divers ainsi qu'à la disponibilité de certaines ressources et outils que le projet de bâtir un « Nouveau » Corpus d'Amsterdam a pris forme. Il s'est concrétisé dans la collaboration étroite des auteurs[2] avec M.-D. Gleßgen (Université de Zurich).

La fiche suivante (tableau 1) présente en sommaire les données les plus importantes du corpus :

titre / référence bibliographique :	*Nouveau Corpus d'Amsterdam*. Corpus informatique de textes littéraires d'ancien français (ca 1150–1350), établi par Anthonij Dees (Amsterdam 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-D. Gleßgen
lieu et date de publication :	Stuttgart : Institut für Linguistik/Romanistik première version publiée en février 2006
site de téléchargement :	http://www.uni-stuttgart.de/lingrom/stein/corpus/ (accès réservé aux chercheurs licenciés)
format	texte avec balisage XML
annotation	parties du discours, catégories flexionnelles, lemmes
époque	environ 1150–1350
taille	299 textes et extraits de textes, 3.184.834 mots
genre	textes littéraires, dont 57 en prose (contenant 29% des mots)

Tableau 1 : Fiche technique du Nouveau Corpus d'Amsterdam

1 Cinq textes ont pu être ajoutés à la collection en 2003 grâce à M. Hans Wesdorp, le gendre de A. Dees.

2 Dans le cadre d'un projet Transcoop III-DEU/1112357 financé entre 2003 et 2006 par l'Alexander von Humboldt-Stiftung, le Conseil de recherches en sciences humaines du Canada (CRSH), et l'Université d'Ottawa (Faculté des Arts).

Le tableau 2 montre la répartition du corpus sur des périodes d'un tiers de siècle, en tenant compte des dates moyennes indiquées par les descripteurs bibliographiques « dateComposition » et « dateManuscrit » :

	date de composition	**date du manuscrit**
période	**nombre de mots**	**nombre de mots**
sans date	429.740	257.341
avant 1133	3.878	0
1133–1166	232.466	7.480
1167–1199	781.328	0
1200–1232	928.158	334.853
1233–1266	287.867	481.839
1267–1299	384.491	732.807
1300–1332	90.751	1.064.945
après 1333	46.155	305.569
total mots	3.184.834	3.184.834

Tableau 2 : nombre de mots par période

1.2 Aperçu historique : l'approche quantitative[3]

En 1971 Anthonij Dees publie l'*Étude sur l'évolution des démonstratifs en ancien et en moyen français* dont onze ans plus tard William Labov (1982:36) dira : « Perhaps the most extensive use of quantitative analysis in historical linguistics is the brilliant monograph of Dees on the development of French demonstrative pronouns and adjectives (1971). » En effet, pour résoudre le problème de la formation des démonstratifs en ancien et en moyen français Dees avait fait appel à une analyse quantitative d'un grand nombre de textes dont la provenance géographique et la date de rédaction étaient connues : les chartes originales, datées et localisées. Il avait découvert ce genre de textes dans le volume III de Schwan-Behrens. Ce volume contient comme on sait 81 chartes du domaine d'oïl choisies de façon représentative. Grâce à cette approche Dees a évité un des plus grands problèmes de l'étude de l'ancien français : le manque de précision des données. De cette façon il a réussi à résoudre l'énigme de l'évolution des démonstratifs. Remarquons qu'en 1831 déjà Gustave Fallot avait signalé l'utilité de chartes pour la recherche linguistique, et que Carolus-Barré en 1964 avait formulé des critères de sélection.

Après lecture de cette étude magistrale, P. van Reenen avait décidé d'aller travailler dans le même style, y voyant le moyen de renouveler les recherches dans le domaine de l'ancien français en général. Il a proposé à Dees de généraliser son approche en commençant par les chartes et d'en faire un corpus destiné à l'exploitation linguistique et dialectale de l'ancien français, domaine d'oïl. Le corpus recueilli par

3 Dans cette partie, nous reproduisons les informations fournies par Pieter van Reenen dans sa communication et dans un fragment qu'il a pu produire avant d'être frappé par la grave maladie qui l'a atteint après l'atelier de Lauterbad.

Dees constituait un bon point de départ. Van Reenen a retenu exclusivement les chartes originales, datées et localisées d'après les critères les plus sûrs, tout en évitant les chartes émanant des autorités dépassant le niveau local, comme les comtes et les évêques, dont il est souvent impossible de déterminer une provenance dialectale précise, d'après les critères de sélection formulés par Carolus-Barré. Vu l'ampleur de la tâche, ces chercheurs ont décidé de sélectionner seulement les chartes déjà publiées, en prenant comme date limite l'année 1300. Une recherche systématique et assidue dans les bibliothèques a permis d'ajouter un bon nombre de chartes utilisables.

En 1976 (voir van Reenen 1976), l'équipe disposait d'un corpus de 3300 chartes locales, originales datées ; suite à une initiative d'informatisation lancée par la Vrije Universiteit Amsterdam, elle a réussi à informatiser des parties importantes des chartes entre 1976 et 1978 : groupes nominaux, groupes verbaux, groupes pronominaux, ordre des mots (en renonçant aux adverbes, conjonctions et noms propres). Le produit le plus important de ce travail a été l'*Atlas des formes et des constructions des chartes françaises du 13^{e} siècle* (Dees et al. 1980).

Pourquoi Dees et P. van Reenen ont-ils réuni ces chartes et localisé par la suite des textes littéraires ? Malgré les avantages que présente l'analyse des chartes pour les études phonologiques et morphologiques, il y a aussi quelques désavantages sur le plan lexical et stylistique (stéréotypes, trop peu de variation ; lexique limité, style formel). Si, en revanche, les désavantages de l'emploi des textes littéraires sont la datation, la localisation et la « Mischsprache », ils présentent les avantages des rimes (permettant de déterminer si oui ou non les voyelles sont identiques, voir Van Reenen (1989)) et d'un lexique riche. Par conséquent, Dees avait complété son étude par l'analyse d'un nombre considérable de textes littéraires, dont le résultat est l'atlas de 1987 (pour la liste complète des publications et les corpus disponibles cf. van Reenen/Schøsler 2000).

Le projet méthodologique de Dees et de P. van Reenen était, en premier lieu, de renouveler les études dans le domaine de l'ancien français, car il n'y avait pas eu de progrès dans les années 60. Tout comme d'ailleurs dans la description des autres langues du moyen âge, ils constataient un certain déséquilibre entre les richesses des données et leur inaccessibilité pour une exploitation systématique. Le malaise de l'état des études de l'ancien français de l'époque était bien illustré par la publication de la *Petite Grammaire de l'ancien picard* en 1949 par Charles Théodore Gossen et le compte rendu exemplaire que Carolus-Barré (1952) a écrit sur cette étude (et qui a d'ailleurs eu un effet salutaire sur son remaniement presque 20 ans plus tard).

Ils avaient constaté une série de problèmes dans les études de l'ancien français, dus au fait que leurs collègues (linguistes et littéraires) n'utilisaient guère les corpus et ne recouraient pas aux moyens de les exploiter avec une précision suffisante pour situer les données dans l'espace et dans le temps. Les deux problèmes majeurs sont les suivants : d'abord celui de la confusion entre les notions d'espace géographique et d'espace dialectal. Par exemple, les domaines (géographiques) de la Picardie et de la France sont limitrophes, tandis que les dialectes picard et francien se recouvrent partiellement. Ainsi, dans son édition du *Chevalier au barisel,* Félix Lecoy localise la langue des manuscrits comme n'étant certainement pas de l'Est et légèrement

teintée de picardismes (1955:XIV–XV). En revanche, la localisation de l'édition d'après l'atlas de Dees et al. (1987:523) indique « Somme/Pas-de-Calais », avec un score de 95. Le second problème est le fait que la langue d'un texte littéraire peut être un mélange ad hoc de dialectes. Il faut alors en calculer le degré de dialecticité.

Les données de l'atlas des chartes ont donc servi à localiser quelques centaines de textes littéraires, qui à leur tour ont fourni le matériel linguistique pour produire l'*Atlas des formes linguistiques des textes littéraires de l'ancien français* (Dees et al. 1987). Pour les textes littéraires, on a défini sur la base de l'atlas des chartes 87 points (qu'on peut réduire à 29), qui fournissent des scores pour le texte littéraire examiné. Quand on compare toujours les traits du texte littéraire avec les traits correspondant dans les chartes et qu'on additionne ces scores, le texte littéraire se localise par définition avec score le plus élevé.[4]

2 RESSOURCES

2.1 Les ressources lexicales

Nous ne présenterons que brièvement les ressources lexicales, leur rôle apparaissant plus clairement dans la partie « Lemmatisation » :

1. La version informatisée du Tobler/Lommatzsch *Altfranzösisches Wörterbuch* (Blumenthal/Stein 2002) contient une liste des 37.000 lemmes de ce dictionnaire ainsi que 15.000 renvois sur les variantes de lemmes.[5]
2. La base de graphies verbales établie par Robert Martin dans les années 60 : cette ressource contient environ 56.600 formes et a ajouté 37.400 formes nouvelles à notre lexique.
3. Les formes graphiques extraites du Godefroy (LFA, Ottawa) : cette ressource contient les lemmes, leurs variantes et les formes fléchies qui apparaissent dans les exemples. Ces 115.000 formes sont annotées pour la partie du discours, mais il n'y aucune information sur les autres catégories morphologiques.
4. Une ressource supplémentaire, compilée manuellement, est le répertoire des mots grammaticaux, particulièrement problématique puisque l'annotation d'origine ne distinguait pas entre certaines catégories qui n'avaient que peu d'intérêt pour les recherches de Dees (cf. la contribution de Y.-C. Morin dans ce volume). Ainsi, l'étiquette « 600 » marque des formes graphiques qui peuvent être adverbes, conjonctions ou pronoms (cf. infra la partie sur la lemmatisation manuelle), et la forme *mais* est toujours annotée en tant que conjonction, malgré ses emplois

4 Note des éditeurs : ce qui paraît conceptuellement simple, ne l'est pas dans l'exécution : les résultats de ces calculs ont été transmis à Hans Goebl, Salzburg, au cours de l'atelier et remplissent une pile considérable de feuilles imprimées. Nous renvoyons également aux informations concernant les scores dans la contribution de Y.-C. Morin dans ce volume.

5 Grâce à l'autorisation de la maison d'édition Franz Steiner, Stuttgart, les listes de lemmes et d'autres matériaux concernant le Tobler/Lommatzsch sont publiquement disponibles : http://www.uni-stuttgart.de/lingrom/stein/tl/

adverbiaux. Ces cas sont gênants pour deux raisons : d'une part, ils devront être désambiguïsés manuellement dans une version ultérieure du corpus ; d'autre part, ils fournissent des séquences d'étiquettes incorrectes et nuisent à l'entraînement de l'étiqueteur. Dans la version actuelle du corpus, nous n'avons pas touché aux étiquettes, mais nous avons systématiquement attribué un lemme à tous les mots grammaticaux du corpus. La source de ces lemmes est marquée par « S » (inspiré par « syntaxe », cf. tableau 3). Pour attribuer ces catégories, on s'est inspiré des conventions CATTEX formulées pour l'annotation de la *Base de français médiéval* (Heiden/Prévost 2005), bien que les abréviations ne soient pas les mêmes. Le résultat est un lexique de 3784 formes attribuées à 135 lemmes pour les étiquettes énumérées dans (1), où le nombre de formes graphiques et un exemple sont indiqués pour chaque catégorie :

(1) ADJ:poss (438, p.ex. mien), CON:coord (23, p.ex. car), DET:def (57, p.ex. li), DET:demo (177, p.ex. cel), DET:ind (646, p.ex. alcun), DET:ndf (79, p.ex. un), DET:poss (401, p.ex. nostre), PRE (448, p.ex. auec), PREDET:a (47, p.ex. al), PREDET:de (41, p.ex. del), PREDET:en (56, p.ex. el), PRO:clit (23, p.ex. en), PRO:demo (341, p.ex. cela), PRO:ind (430, p.ex. alcun), PRO:invar (247, p.ex. quoi), PRO:pers (310, p.ex. el), PRO:poss (238, p.ex. mien).

2.2 Unifier les ressources

Le tableau 3 résume les ressources lexicales et textuelles indiquant le code utilisé pour marquer la source du lemme dans l'annotation, le nombre des formes graphiques, et les informations fournies par la ressource :

Ressource	**code**	**formes graphiques**	**informations**
Tobler/Lommatzsch	T	58727	lemme, variantes
graphies verbales (Martin)	M	71265	catégorie, lemme
graphies du Godefroy	G	115498	catégorie, lemme
mots grammaticaux	S	3999	catégorie, lemme
lemmatisation manuelle	Z	4456	catégorie, lemme
index	I	41377	catégorie, lemme
chartes de l'Aube	C	4260	catégorie, lemme
Amsterdam Corpus	A	133894	catégorie

Tableau 3 : Ressources lexicales

On a fusionné ces ressources dans un dictionnaire des formes d'ancien français en convertissant l'information morphologique à un format généralisé, produisant ainsi un inventaire de 50 étiquettes effectivement utilisées pour entraîner l'étiqueteur. Ces étiquettes distinguent entre la partie du discours et certaines sous-catégories, par exemple adjectifs numéraux ou pronoms (indéfinis, interrogatifs). Bien que la plupart de nos ressources fournissent des catégories plus fines, nous avons opté pour

ce jeu d'étiquettes réduit : d'une part, parce qu'il facilitait l'intégration des indices et des formes extraites du Godefroy (avec des catégories pauvres en traits morphologiques) ; de l'autre, parce que le jeu d'étiquettes réduit nous a permis de désambiguïser les formes homographes avec une meilleure précision et par conséquent d'attribuer le lemme correct avec une précision supérieure.

Après élimination des entrées doubles, ce lexique uni contient 235.000 formes graphiques différentes. L'unification des ressources ne s'effectue pas au détriment de l'information sur les lemmes : si les lemmes diffèrent pour une forme graphique donnée, ils sont répertoriés comme alternatives dans le lexique. Les majuscules listées dans le tableau 3 sont reliées aux lemmes par un trait de soulignement pour indiquer la source du lemme (cf. tableau 4) et apparaîtront plus tard dans l'annotation du corpus (cf. infra les commentaires sur l'attribut « src », et le tableau 10).

Finalement, un programme appliquant des règles de variation orthographique opérant sur le lexique unifié propose un lemme pour environ 127.000 formes non lemmatisées (donc des formes qui n'ont aucune correspondance lemmatisée dans une de nos ressources). Pour la version actuelle, environ cinquante règles s'appliquent à certaines chaînes de caractères, surtout à la fin du mot, les remplacent par une variante graphique et vérifient la présence d'un lemme pour cette forme modifiée. Certaines de ces règles portent des restrictions liées à la partie du discours. S'il existe un lemme (pour une forme de la même catégorie, évidemment), il est associé à la forme d'origine, marqué par un astérisque (qui sera conservé dans l'annotation du corpus) et par un code identifiant la règle appliquée dans le lexique. Nous nous sommes bornés à formuler des règles prudentes, et nos expériences ont montré que le résultat est très fiable. Dans l'exemple (2), la règle de correspondance « u[mn]-o[mn] » a été appliquée pour attribuer le lemme *abondance* à la forme *abundance* :

(2) abundance NOM *{abondance;u[mn]-o[mn]}_Alabondance_T

La dernière étape est la comparaison de chaque lemme du lexique avec la liste des lemmes du Tobler/Lommatzsch qui nous sert de référence: chaque lemme répertorié dans le Tobler/Lommatzsch est marqué par un plus, par exemple « +abiter_I ».

forme	**cat1**	**lemme1**	**cat2**	**lemme2**
abitablement	NIL	habitablement_G		
abitacion	NOM	+abitacion_T		
abitacle	NOM	+abitacle_T		
abitance	NIL	habitance_G	NOM	+abitance_T
abitant	NOM	+abitant_T	VER	*+abiter_IT
abitanz	NOM	*+abitant_T		
abitanze	NIL	habitance_G		
abitast	VER	<nolem>		
abitateur	NIL	habitateur_G		
abitation	NOM	*+abitacion_T		
abitations	NOM	*+abitacion_T		

forme	**cat1**	**lemme1**	**cat2**	**lemme2**
abitator	NOM	+abitator_T		
abite	VER	+abiter_I		
abitee	NIL	habiter_G	VER	*+abiter_IT

Tableau 4 : Le lexique unifié des formes d'ancien français

Le tableau 4 montre à titre d'exemple quelques entrées du lexique unifié. La première colonne contient la forme graphique, les paires de colonnes suivantes indiquent la catégorie et le lemme. Les formes ambigues (*abitant* par exemple) ont plus d'une paire catégorie-lemme. Certaines formes comme *abitablement* n'ont pas de catégorie (donc « NIL »), d'autres comme *abitast* n'ont pas de lemme. Pour la forme *abitance,* le lexique fournit deux lemmes différents (tirés du Godefroy et du Tobler/Lommatzsch, respectivement), et pour *abitanz, abitation(s),* et *abitee,* les règles présentées ci-dessus ont proposé un nouveau lemme (*), qui a pu être vérifié dans le Tobler/Lommatzsch (+).

3 LEMMATISATION

Pour les grands corpus électroniques, nous considérons que la catégorisation grammaticale et la lemmatisation ne relèvent pas du luxe, mais au contraire constituent des opérations essentielles. Procédures dangereuses parfois, qui peuvent gauchir ou fausser les données (d'où la nécessité d'offrir en parallèle des données brutes, un index brut, par exemple, à côté de l'index lemmatisé), mais elles s'imposent néanmoins si on veut exploiter finement les corpus et en tirer des observations statistiques intéressantes.

Le corpus établi à Amsterdam était, comme nous l'avons indiqué, largement indexé grammaticalement, de façon souvent très précise malgré la fameuse catégorie 600, déjà mentionnée, fourre-tout où l'on trouve pêle-mêle tous les *ou*, *ne*, *si*, *qui* et *que* avec pas moins de 226.329 occurrences (ce que nous avons rebaptisé PROCON, terme vague à souhait : pronom/conjonction...). Certaines graphies du corpus étaient même désambiguïsées à l'aide de numéros ; il importait donc d'y ajouter un effort de lemmatisation.

Celle-ci s'est faite en plusieurs étapes : essentiellement par l'apport de ressources externes (des listes de lemmes couplés aux formes graphiques qu'ils subsument) et par une lemmatisation manuelle, d'une part, mais aussi, d'autre part, par l'attribution automatisée de lemmes probables. Nous avons déjà parlé de ce second cas ; nous nous attacherons maintenant au premier.

Les ressources externes qui ont été mises à contribution sont celles que nous regroupons dans notre projet SOFA (Sources et Outils pour l'Étude du Français Médiéval), dirigé par A. Stein, M.-D. Gleßgen et P. Kunstmann. Ce projet vise, entre autres, à constituer un dictionnaire électronique des formes graphiques occurrentes. A. Stein terminait la version électronique (format image) du *Tobler-Lommatzsch* ; M.-D. Gleßgen travaillait à l'élaboration d'un corpus électronique de

chartes ; de son côté, P. Kunstmann préparait une *Base Lemmatisée d'Ancien Français*, dont il avait présenté les principes constitutifs en 2001 à Salamanque, au congrès de la Société de Linguistique Romane. Nous avons décidé, tous les trois, d'unir nos efforts pour constituer le dictionnaire électronique en question.

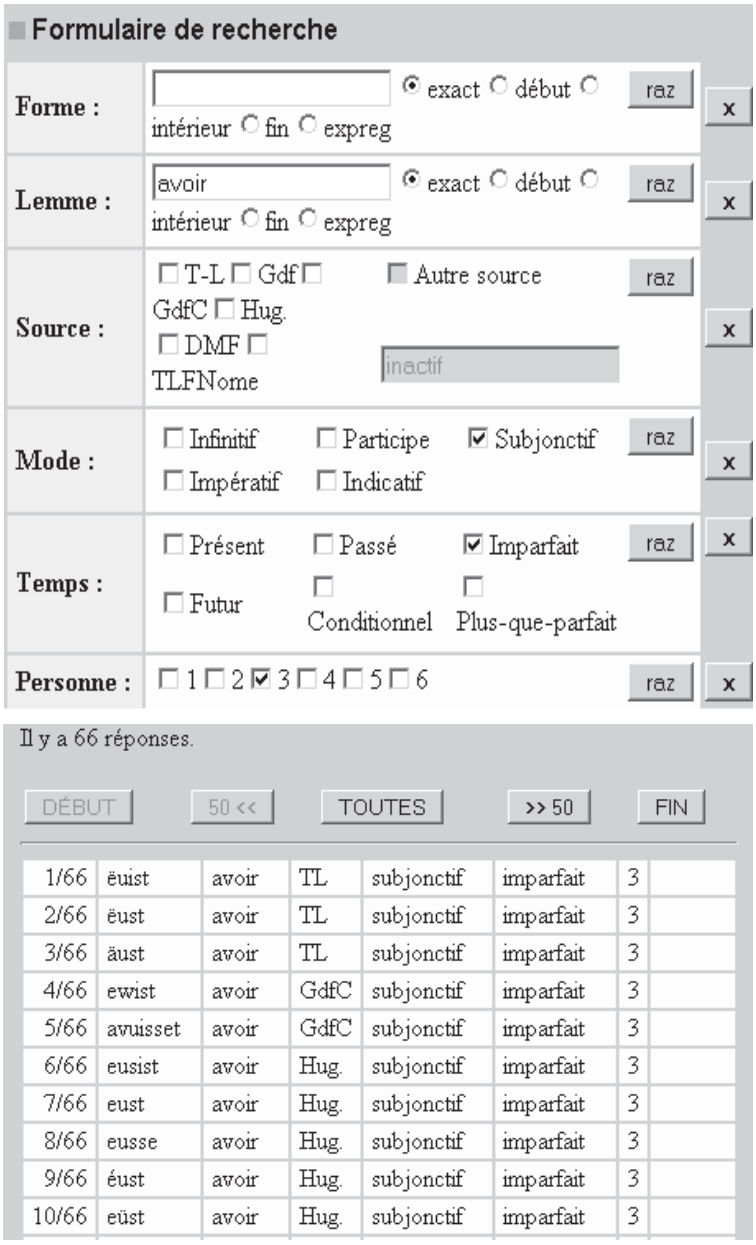
G. Roques avait suggéré à P. Kunstmann l'idée d'une base lemmatisée. Index et concordances offraient, certes, des matériaux considérables ; mais comment les exploiter utilement sans regrouper les formes systématiquement et sous des vedettes convenues ? Il lui avait conseillé, à cet effet, de saisir les graphies du *Godefroy* pour faciliter le travail de lemmatisation.

P. Kunstmann a ainsi fait saisir, à Ottawa, une bonne partie des graphies du *Godefroy*, avec leur lemme correspondant. À partir d'un certain moment, son équipe a laissé tomber les verbes pour se concentrer sur les autres parties du discours. C'est qu'elle avait obtenu une autre source, bien plus riche, de graphies verbales.R. Martin avait, en effet, proposé en 1994 d'envoyer à Ottawa, dans le cadre d'un accord de collaboration scientifique entre l'INALF (Institut National de la Langue Française, auquel s'est substitué maintenant l'ATILF –Analyse et Traitement Automatique de la Langue Française- à Nancy) et le LFA (Laboratoire de Français Ancien, Université d'Ottawa), l'immense fichier de graphies verbales (autour de 20.000 fiches) qu'il avait constitué dans les années soixante à la demande de P. Imbs. Le fichier a été principalement établi à partir des trois dictionnaires de Tobler-Lommatzsch, Godefroy et Huguet, de neuf livres (études, monographies, manuels) de morphologie et de grammaire, mais aussi d'éditions de texte, de glossaires et d'articles spécifiques. Le lemme retenu pour chaque fiche (la forme d'infinitif) correspond à celui de la source utilisée ; les formes graphiques sont classées suivant le mode (non personnel dans le tiers supérieur de la page, personnel dans les deux-tiers inférieurs), le temps et la personne.

Ces fiches ont été saisies sur logiciel FileMaker à Ottawa, de 1995 à 2000, à l'exception d'un bon nombre de formes verbales préfixées ; par exemple, a été saisi *dire*, mais pas *interdire* ; toutefois on trouve *contredire* et *redire*... Retour à l'envoyeur : le contenu du fichier se trouve maintenant en grande partie à l'ATILF ; c'est la *Base de graphies verbales de l'Ancien Français à la Renaissance*, dont G. Souvay a établi en 2004 un prototype.

Chaque fiche comporte huit rubriques (tableau 5) : forme ; lemme ; dictionnaires ; autres sources ; mode ; temps ; personne ; remarques. La base peut être interrogée à partir d'une seule rubrique ou de plusieurs à la fois. La publication sur le site se fera progressivement dans l'ordre alphabétique des lemmes. Les graphies comportant un préfixe avaient été regroupées à l'INaLF sous leur racine ; elles ont été dégroupées au LFA et placées alphabétiquement d'après leur lettre initiale. En conséquence, la base ne sera véritablement complète, même pour les premières lettres, que lorsque l'ensemble du travail sera achevé.

À côté du *Godefroy* et de cette base, nous avons utilisé divers index lemmatisés établis au LFA, du plus ancien (celui du *Couronnement de Louis*, 1997) au plus récent (celui de la *Chanson de Roland*, 2003). Pour le premier, les lemmes étaient empruntés au *Tobler-Lommatzsch* ou, à défaut, au *Godefroy* ; pour les lemmes grammaticaux n'étaient donnés que les deux premières références ; les noms propres



1/66	ëuist	avoir	TL	subjonctif	imparfait	3	
2/66	ëust	avoir	TL	subjonctif	imparfait	3	
3/66	äust	avoir	TL	subjonctif	imparfait	3	
4/66	ewist	avoir	GdfC	subjonctif	imparfait	3	
5/66	avuisset	avoir	GdfC	subjonctif	imparfait	3	
6/66	eusist	avoir	Hug.	subjonctif	imparfait	3	
7/66	eust	avoir	Hug.	subjonctif	imparfait	3	
8/66	eusse	avoir	Hug.	subjonctif	imparfait	3	
9/66	éust	avoir	Hug.	subjonctif	imparfait	3	
10/66	eüst	avoir	Hug.	subjonctif	imparfait	3	

Tableau 5 : Exemple de requêtes et premiers résultats

n'étaient pas traités. Pour le dernier, toutes les références sont données ; on a tenu compte de tous les noms propres et on a utilisé, pour distinguer les lemmes homonymes, les indices numériques retenus par les auteurs du DVD du *Tobler-Lommatzsch* ; ce qui a conduit d'ailleurs à un étiquetage de tous les mots du texte, dont voici le début :

```
1. Carles_<Carle-sp> li_<le-art> reis_<roi2-sm>, nostre_<nostre-
   pron/adjposs/sm> emperere_<emperëor-sm> magnes_<maigne-adj>,
```

```
2. Set_<set-card> anz_<an-sm> tuz_<tot-adj/pron/adv> pleins_<plein-
   adj> ad_<avoir-v> estet_<estre1-v> en_<en1-prép> Espaigne_<Es-
   paigne-sp>
3. Tresqu'_<tresque1-prép/conj> en_<en1-prép> la_<le-art|le-art>
   mer_<mer-s> cunquist_<conquerre-v> la_<le-art|le-art> tere_
   <terre-sf> altaigne_<hautain-a.>.
```

Les autres index lemmatisés du LFA concernent Chrétien de Troyes : ce sont ceux du *Chevalier au Lion* et du *Conte du graal* ; ils ont été modifiés sur le modèle de celui du *Roland* et ont donné lieu à un étiquetage des romans, base indispensable pour le projet *DÉCT* (*Dictionnaire Électronique de Chrétien de Troyes*). A également été saisi le *Lexique* de Chrétien de Troyes par M.-L. Ollier, accompagnant sa concordance parue en 1986 ; nous en avons modifié certains lemmes et certains indices grammaticaux pour l'adapter à notre système de lemmatisation ; nous avons notamment introduit les indices numériques de désambiguïsation des lemmes du *Tobler-Lommatzsch* sur DVD. Voici quelques lignes à titre d'exemple :

lit1	sm	lit liz
litiere	sf	litiere
livre2	sm	livre livres
livre3	s	livre livres
livrer	v	liverra liverroient livrassent livre livré livree livrees livrent livrer livrera

À cela se sont ajoutés les lemmes du corpus des chartes de l'Aube, que nous a communiqué récemment P. van Reenen (cf. pour les chartres Van Reenen (2007)).

Le DVD du *Tobler-Lommatzsch* constitue notre dernière source externe de lemmes, et une source considérable, comme il a été mentionné plus haut ; à vrai dire, ces lemmes ne sont pas accompagnés de graphies, mais ils se présentent souvent avec une série de variantes et constituent ainsi un apport précieux pour notre entreprise.

Si le *Tobler-Lommatzsch* représente par excellence, ou tout au moins par convention, notre répertoire de lemmes et si nous y adhérons étroitement, il arrive exceptionnellement que nous nous en écartions. Pour les lemmes lexicaux, la seule exception concerne les adverbes de manière se terminant par le suffixe *-ment*. Ils sont traités, dans le dictionnaire, sous l'entrée du mot dont ils sont dérivés ; nous les en extrayons, au contraire, et leur conférons le plein statut de lemme. Ainsi, même dans le DVD, il n'est pas très facile de trouver l'adverbe *grandement*, contrairement à la forme de comparatif *graignor*, considérée comme un lemme, qui a droit à un article particulier. Si l'on cherche sous l'adjectif *grant*, l'adverbe n'apparaît même pas parmi les renvois ; il faut cliquer sur *grant* et ouvrir l'article, pour l'y trouver relégué à la fin. Donner à l'adverbe le statut de lemme (*granment*) permet à Twic de fournir la liste des occurrences des diverses formes (tableau 6).

```
   search amslit for: lemma=granment
   <hit hitnr=1> qui en mer entre s a granment
<hit hitnr=3> grantment fu garis li enfes et
<hit hitnr=12> tant gramment que tuit li plusor
```

```
<hit hitnr=17> granmant que son non ne seusse
<hit hitnr=18> puis ne demora pas graument
<hit hitnr=19> et puis ne demoura mie gramment apres que li cuens
thiebaus morut , si laissa yy livres as croisiés et a chelui qui
apres lui seroit kievetaine et sires des croisiés , et a metre la
ou li croisié vauroient .
<hit hitnr=25> s amor en creistra grandement
<hit hitnr=30> grantmant mie ne mesprenoient
<hit hitnr=31> ancoisqu i ussent granmant sis
```

Tableau 6 : Recherche Twic

C'est essentiellement avec les lemmes grammaticaux que nous prenons quelque distance par rapport au *Tobler-Lommatzsch* ; le terrain est piégé : le choix du lemme est parfois discutable, l'analyse grammaticale souvent contestable. C'est en particulier le cas de certains déterminants du substantif, pour lesquels des corrections ont été nécessaires. Ainsi on est un peu surpris que l'adjectif indéfini distibutif *chasque* soit analysé comme « pron. indef. » dans l'en-tête de l'article du *Tobler-Lommatzsch*, alors que dans le corps de l'article on précise bien qu'il s'agit d'un adjectif et que tous les exemples cités relèvent de ce seul emploi. Dans les index du LFA, on l'a considéré comme adjectif ; dans le Corpus d'Amsterdam il a été analysé comme déterminant indéfini.

À voir les quatre lemmes démonstratifs du *Tobler-Lommatzsch*, on pourrait croire qu'il ne s'agit que de pronoms :

ce1 :	pron. demonstr. neutr.	cen, ceu, cié, ço, çou
ce2 :	pron. demonstr.	
cel :	pron. demonstr.	cil, celui, cel, cil, ceus, cele, celi, celes, celor
cest :	pron. demonstr.	

Ce n'est en fait le cas que pour le premier, *ce1*. *Ce2* n'est pas pronom mais adjectif : l'article est composé d'un seul paragraphe, qui commence par l'indication « adj. masc. dieser. jener ». Quant à *cel* et *cest*, ils peuvent être, bien sûr, adjectifs aussi bien que pronoms. Ce que P. Kunstmann précise ainsi dans l'index de Chrétien :

ce1	pron dém neutre	c' ç' ce çué ceu
ce2	pron/adj dém	ce ces
cel	adj dém	çaus çax ce cel cele celes celi cels celui ces cil
cest	pron/adj dém	ce ces cest ceste cestes cesti cestui cez cist

L'équipe d'Amsterdam avait indexé minutieusement les formes de *ce1* sous la catégorie « PRO:invar » ; la liste en est longue ; TWIC nous présente :

ce1 :	aiso, ce, cen, ceo, ceu, ch, che, chem, chou, chu, co, cou, cu, eco, se, seci, zo

C'est surtout dans le cas des possessifs que les modifications s'imposent : ainsi *lor2* reçoit, dans *Tobler-Lommatzsch*, la catégorisation « pron. possess. », tandis que le corps de l'article présente d'abord l'emploi adjectival (« ihr »), qui a la part du lion

par rapport à l'emploi substantival (*le lor* « das Ihrige ») qui suit. Les index du LFA le présentent comme suit : « *lor2* pron/adj poss ». On trouve, dans le Corpus d'Amsterdam, trois catégories : « ADJ:poss », « DET:poss », « PRO:poss » ; mais les deux premières, référant au même emploi, devraient être regroupées. Quant aux possessifs de l'unité, la présentation est encore plus confuse. *Godefroy* distingue formes atone et tonique, mais sans le dire explicitement ; les codes grammaticaux sont plus ou moins semblables, ainsi que les définitions :

2. mon	adj. poss.	qui est à moi
mien	adj. poss.	qui est à moi
1. ton	adj. poss.	qui est à la personne à qui l'on parle
tien	adj. masc.	qui est à la personne à qui on parle
2. son	adj. poss.	qui est à la personne ou qui dépend de la chose dont on parle
sien	adj.	qui est à la personne dont on parle

Le tableau du *Tobler-Lommatzsch* (les lemmes sont cités ici avec les indices de la version électronique) est remarquable par sa dissymétrie :

Mon1, ma (m'), mes, mi *pron. poss.*	Mien , *fem.* moie *pron. poss.*
Ton4 *pron. poss. masc. 2 Person (unbetont)*	
	Suen, suon, sien, son *pron. poss 3 pers masc. u. a.; siene, siue, söe, soie, sa u. a. pron. poss. 3 pers. fem.*

Formes atone et tonique pour la première personne, atone pour la deuxième, tonique pour la troisième ; toutes ces formes sont tenues pour des « pron. poss.», la valeur adjectivale n'étant nulle part mentionnée dans l'en-tête de l'article.

Dans son index de Chrétien, P. Kunstmann a retenu pour chaque personne un seul lemme (qui correspond à la forme atone), analysé comme « pron/adj poss ». Ce qui l'a conduit à créer le lemme *son4*.

mon1	pron/adj poss	m' ma mes mi mon mien miens moie moies
ton4	pron/adj poss	t' ta tes toe ton tuen tuens
son4	pron/adj poss	s sa se ses si son soe soes suen suens

Cette présentation a été adoptée par A. Stein dans le corpus Twic :

3425	mon	DET:poss:obj:masc:sg	mon1
277	mien	ADJ:poss:obj:masc:sg	mon1
2	ton	ADJ:poss:obj:masc:sg	ton4
120	tuen	ADJ:poss:obj:masc:sg	ton4
34	son	ADJ:poss:obj:masc:sg	son4
75	suen	PRO:poss:obj:masc:sg	son4

Piégé au départ, le terrain des mots grammaticaux reste miné dans l'état actuel de notre outil ; les bogues sont encore nombreux ; nous recommandons aux utilisateurs

d'être particulièrement prudents. Cette réserve, honnête, n'enlève cependant rien à l'utilité certaine du programme dans ce domaine.

Une fois toutes ces sources de graphies et de lemmes engrangées dans Twic et utilisées par le logiciel, il restait cependant un grand nombre de formes dont le lemme n'était pas identifié. Force nous a été de procéder à une lemmatisation manuelle, effectuée par P. Kunstmann à Ottawa, par parties du discours, en commençant par les graphies à fréquence élevée. Les formes verbales, les adjectifs et les adverbes ont été lemmatisés jusqu'à la fréquence 3 incluse ; les formes nominales jusqu'à la fréquence 10 incluse ; les noms propres (lemmatisés sous la forme du cas régime quand c'était possible) jusqu'à 10 également ; les conjonctions intégralement. Pour les homographes, deux ou plusieurs lemmes ont été proposés. Citons l'exemple d'*esforz1* et *2* :

16	effors	NOM:obj:masc:sg	esforz1	esforz2
20	effort	NOM:obj:masc:sg	esforz1	esforz2

qui correspondent aux deux vedettes suivantes du *Tobler-Lommatzsch* :

– *esforz1* sm [*FEW* III 727b *fortia*] « Heer »
– *esforz2* sm [*FEW* III **fortiare*] « Anstrengung ».

Tâche ingrate certes, mais indispensable – en particulier pour les textes transcrits directement de manuscrits : les auteurs des transcriptions ont laissé à dessein, par souci de fidélité au manuscrit, *i* et *u*, sans recourir aux lettres raméennes *j* et *v*. Comme les formes sans lemme provenaient surtout de transcriptions, le travail de lemmatisation était somme toute assez commode.

4 CONSTRUCTION DU CORPUS

4.1 Étapes techniques

Le corpus d'Amsterdam contient environ 200 textes et extraits de textes, certains en plusieurs versions (manuscrits), ce qui donne un total de 299 fichiers contenant 3.184.834 mots (*tokens*). L'équipe de Dees avait annoté ces formes manuellement en utilisant un jeu de 225 étiquettes numériques représentant en trois chiffres la partie du discours et d'autres traits morphologiques (p. ex. « 566 » pour verbe, futur, sixième personne) ; le tableau 7 reproduit quelques lignes du fichier original du *roman de Renart*. En revanche, les textes ne contiennent pas de ponctuation, et les chiffres et noms propres ont parfois été omis ou abrégés.

```
c6/ren2: or_311 me_412 *covient_513 tel_026 chose_006 dire_592
c6/ren2: dont_341 je_411 vos_451 *puisse_521 faire_592 rire_592
c6/ren2: car_331 je_411 *sai_511 bien_311 ce_341 *est_513 la_105 pure_025
c6/ren2: que_600 de_301 sarmon_002 n_319 *avez_515 vos_451 cure_006
c6/ren2: ne_600 de_301 corsaint_002 oir_592 la_106 vie_006
```

```
c6/ren2: de_301 ce_341 ne_319 vos_451 *prant_513 nule_185 envie_005
c6/ren2: mais_331 de_301 tel_026 chose_006 qui_600 vos_451 *plaise_523
c6/ren2: or_311 *gart_523 chascuns_281 que_600 il_431 se_600 *taisse_523
c6/ren2: que_600 de_301 2bien_002 dire_592 *sui_511 en_319 voie_006
c6/ren2: et_331 toz_281 garniz_581 se_600 diex_001 me_412 *voie_523
```

Tableau 7 : Format original du Corpus d'Amsterdam

Le tableau 8 visualise le traitement des ressources et l'élaboration du Nouveau Corpus d'Amsterdam. Le principe central de la mise en oeuvre était de préserver l'information originale du projet de Dees et par conséquent d'ajouter de nouvelles informations d'une manière transparente et reproduisible. Pour atteindre cet objectif, toute nouvelle version du corpus (par exemple après amélioration des ressources lexicales ou bibliographiques) est générée en appliquant les outils aux fichiers d'origine. Dans ce qui suit, nous décrirons les étapes successives de ce traitement.

Dans la section 2.2 nous avons décrit l'unification des ressources lexicales. Dans les étapes suivantes (visualisées dans le cadre « Tagging », tableau 8), le lexique unifié est utilisé par l'étiqueteur pour produire les paramètres, un fichier binaire contenant les entrées du lexique ainsi que les probabilités lexicales et contextuelles (cf. section 4.2 pour plus de détails sur le TreeTagger).

Le processus de reformatage des fichiers originaux est symbolisé dans le cadre « XML Mark-up ». Pour chacun des textes, les informations répertoriées dans la bibliographie sont projetées dans la balise XML <subcorpus> (cf. la section suivante). Ensuite, le texte est segmenté en mots (balise <word>) tout en conservant l'annotation originale dans l'attribut « deespos ». Les mots sont traités à part par le

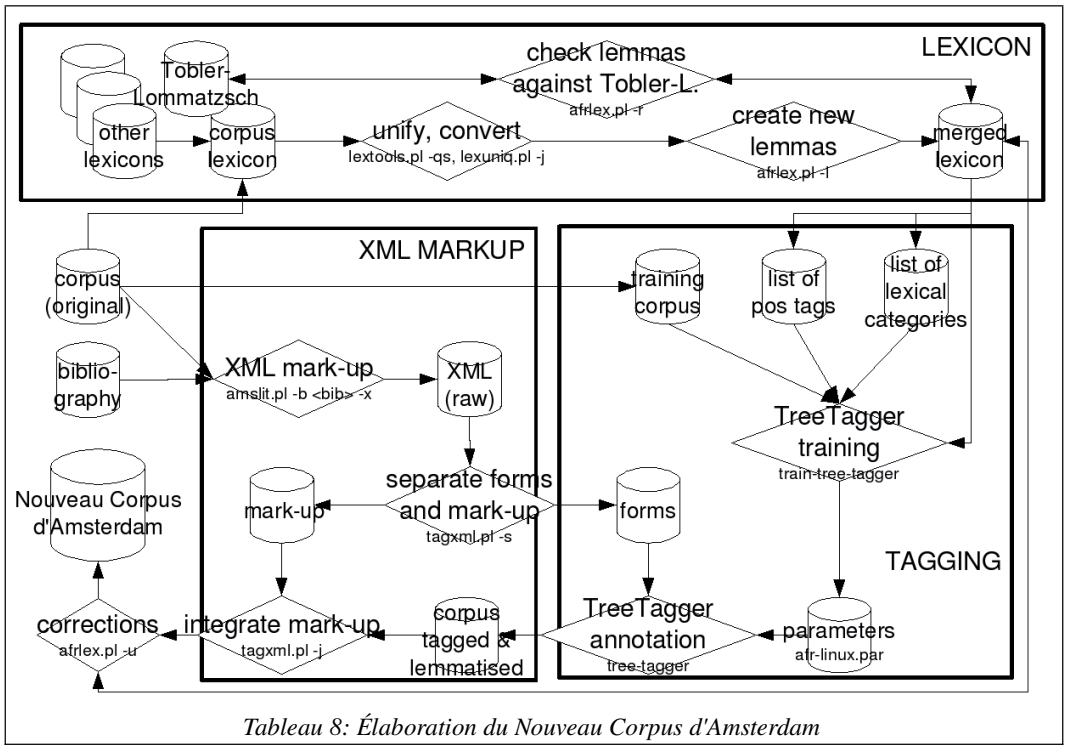


Tableau 8: Élaboration du Nouveau Corpus d'Amsterdam

TreeTagger, pour y ajouter les étiquettes du lexique et le(s) lemme(s) associé(s) à la forme graphique en utilisant les paramètres pour distinguer entre formes homographes. A la sortie de l'étiquetage, le balisage XML est réintroduit dans le texte. Enfin, le lexique est réutilisé pour effectuer des corrections finales (p. ex. traitement de certaines différences entre l'étiquetage manuel et les étiquettes proposées par le TreeTagger).

4.2 Les outils: l'étiquetage automatique

L'étiqueteur « TreeTagger » qui a servi à intégrer les lemmes dans nos textes, est également un outil précieux pour étiqueter et lemmatiser des textes nouveaux, puisque le fichier des paramètres résultant de l'entraînement sur le corpus d'Amsterdam est librement disponible. Nous expliquons donc brièvement le fonctionnement de l'étiqueteur (pour plus de détails, cf. Schmid 1994 et Stein/Schmid 1995).

TreeTagger estime les probabilités de transition à l'aide d'un arbre de décisions binaires. La probabilité d'un trigramme (une série de trois étiquettes) est déterminée en parcourant l'arbre jusqu'à la feuille. Ainsi, pour calculer la probabilité du trigramme DET-ADJ-NOM, il faut d'abord vérifier, en partant de NOM, si le mot précédent est un ADJ. Si oui, le test suivant porte sur DET, situé en première position du trigramme, et aboutit à la feuille, qui contient les probabilités des différentes étiquettes.

Cet arbre de décisions est construit de façon récursive à partir d'un jeu de trigrammes extrait du corpus d'entraînement (ici l'annotation originale du corpus d'Amsterdam) : à chaque pas récursif, un test est formulé qui divise le jeu de trigrammes en deux sous-jeux qui sont le plus distincts possible par rapport à la distribution de la dernière (ici la troisième) étiquette. Ce test vérifie si une des étiquettes précédentes est égale à une étiquette donnée. A chaque pas récursif, tous les tests possibles sont comparés, et c'est celui avec la plus grande valeur informative qui est rattaché au noeud. Après élimination des feuilles apportant un gain d'information insuffisant (le noeud mère devient alors une feuille), l'arbre utilisé pour notre corpus ne contient plus que 540 noeuds (comparé à 1700 avant « élagage »).

Puisqu'un dictionnaire ne peut être exhaustif, TreeTagger crée des ressources supplémentaires pour résoudre le problème des mots inconnus : le dictionnaire des suffixes et le dictionnaire des préfixes. A partir de tous les suffixes d'une longueur *n* des classes ouvertes (donc productives), TreeTagger construit un arbre de lettres. Les probabilités $p(e|s)$ des étiquettes *e* en cooccurrence avec un suffixe *s* sont rattachées aux feuilles de l'arbre. Le dictionnaire des préfixes est construit de façon analogue. TreeTagger recourt aux dictionnaires d'affixes seulement si la recherche dans le dictionnaire des formes fléchies (ici : le dictionnaire des graphies d'ancien français unifié) a échoué. Il proposera donc une étiquette probable pour ces formes, mais ne pourra évidemment pas proposer un lemme.

Le fonctionnement autonome de TreeTagger permet d'annoter de grands corpus rapidement de façon non surveillée. D'autre part, les éditeurs de textes anciens (d'une ampleur réduite) pourront utiliser les options produisant un étiquetage ambigu, indiquant pour les homographes toutes les étiquettes possibles avec leurs

probabilités respectives. TreeTagger indiquera également si les étiquettes proviennent du lexique ou si elles ont été « devinées » à partir d'une analyse des préfixes ou des suffixes.

La performance de TreeTagger a été évaluée sur les 500.000 derniers mots du corpus ; par conséquent, 2,7 millions de mots du corpus constituaient le corpus d'entraînement. La précision de l'étiquetage est de 92,7% (relativement modeste comparée aux 98,5% pour le français moderne). Le tableau 9 présente les taux de lemmatisation, classifiés par partie du discours :

POS	**types**		**tokens**	
ADJ	2129/2774	76.75%	24009/24795	96.83%
ADV	1012/1285	78.75%	36514/36923	98.89%
CON	136/ 136	100.00%	26497/26497	100.00%
DET	264/ 264	100.00%	49427/49427	100.00%
GDF	8/8	100.00%	16/ 16	100.00%
INT	16/ 28	57.14%	190/ 251	75.70%
NOM	6915/9483	72.92%	76113/80050	95.08%
NPR	657/2185	30.07%	6206/ 8800	70.52%
PON	11/ 11	100.00%	14212/14212	100.00%
PRE	177/ 177	100.00%	42914/42914	100.00%
PREDET	29/ 29	100.00%	6838/ 6838	100.00%
PRO	438/ 438	100.00%	79774/79774	100.00%
PROCON	18/ 18	100.00%	21832/21832	100.00%
VER	14336/7282	82.95%	104439/07652	97.02%
total		**76.60%**		**97.80%**

Tableau 9 : Taux de lemmatisation

4.3 Balisage XML

Le balisage XML de cette première version du Nouveau Corpus d'Amsterdam (2006) est assez rudimentaire et a été conçu pour satisfaire les exigences de deux programmes de recherche, Xaira (Oxford University Computing Service) et Twic (ILR, Université de Stuttgart), pour lesquels la première distribution du corpus contenait des versions formatées.

```
<subcorpus id=«abe» deaf=«JMeunAbB» titreDees=«J. de Meun, Traduction de
la première epître de P. Abélard, v. 1-821» editionDees=«éd. Ch. Char-
rier, Paris 1934» manuscritDees=«Paris, Bibl. Nat., fr. 920»
regionDees=«REGION PARISIENNE» codeRegional=«54» coefficientRegional=«84»
vers=«non» ponctuation=«non» mots=«18183» passage=«v. 1-821/1588» comment
airePhilologique=«éd. C. CHARRIER, ms. BN fr. 920» qualite=«ms3»
commentaireForme=«nil» auteur=«JEAN DE MEUN» dateComposition=«1280» date-
Manuscrit=«1395» lieuComposition=«frc.» lieuManuscrit=«Paris» genre=«nil»
traditionTextuelle=«nil» analyses=«nil»>
<s line=«1»>
```

```
<word pos=«NOM:suj:masc:pl» deespos=«003» lemma=«essemple»
src=«+I»>essamples</word>
<word pos=«VER:ppre:pl» deespos=«586» lemma=«UNKNOWN»>attaignens</word>
<word pos=«PROCON» deespos=«600» lemma=«en1+le» src=«S»>ou</word>
<word pos=«VER:ppre:pl» deespos=«586» lemma=«UNKNOWN»>appaissans</word>
<word pos=»ADV» deespos=«311» lemma=«sovent» src=«*IT»>souvent</word>
<word pos=«DET:def:obj:masc:pl» deespos=«104» lemma=«le» src=«S»>les</
word>
<word pos=«NOM:obj:masc:pl» deespos=«004» lemma=«talent»
src=»*+IT»>talens</word>
<word pos=«PREDET:de:obj:masc:pl» deespos=«114» lemma=«de+le»
src=«S»>des</word>
[. . .]
```

Tableau 10 : Balisage du Nouveau Corpus d'Amsterdam

Au niveau du texte, la balise <subcorpus> délimite les différents fichiers du corpus original et contient les paires attribut-valeur pour les descripteurs de l'entrée bibliographique correspondante. Bien que tous les descripteurs ne soient pas encore complets (p. ex. genre= « nil »), le travail bibliographique a été intensifié depuis la publication de la première version, et le tableau 10 montre les descripteurs plus complets de la deuxième version prévue pour 2007 avec des informations de dates (de composition : 1280, du manuscrit : 1395) et de qualité (de la transcription ou de l'édition critique : qualite= « ms3 », indiquant que la transcription du manuscrit est d'une qualité moindre). M.-D. Gleßgen et al. (dans ce volume) fournissent une description détaillée des aspects philologiques de cette classification qui, évidemment, est cruciale pour un corpus médiéval. La classification diatopique des textes qui présente le résultat principal du travail de Dees est répertoriée dans les attributs « regionDees » (p. ex. « région parisienne »), « codeRegionalDees » (p. ex. « 54 », qui est synonyme de « région parisienne »), et « coefficientRegionalDees » (p. ex. « 84 » sur une échelle exprimant l'affinité avec une région avec un maximum théorique de 100). Les attributs diatopiques introduits par M.-D. Gleßgen indiquent parfois une localisation alternative des textes (« lieuComposition ») et des manuscrits (« lieuManuscrit »). Enfin, l'attribut « deaf » relie le texte à la bibliographie exhaustive du *Dictionnaire étymologique de l'ancien français* (DEAF), également disponible en ligne.[6]

La représentation non structurée de l'information bibliographique à l'intérieur de la balise <subcorpus> a été préférée à d'autres formes de représentation, comme par exemple un en-tête séparé dans le style de la Text Encoding Initiative (TEI), puisqu'elle permettait facilement de définir des subcorpus pour certaines régions ou époques en appliquant des expressions régulières à ces attributs. Puisque Xaira et Twic permettent également de projeter les attributs dans les occurrences affichées lors d'une recherche, il est facile de classer le résultat d'après la date, la région, etc.

Dans un corpus traditionnel, le niveau inférieur au texte est en général le paragraphe ou la phrase. Les fichiers originaux du corpus d'Amsterdam ne contenaient

6 Les noms d'attributs correspondent à la version 2 du corpus, prévue pour 2007.

aucune structure de ce type, hormis le retour de chariot encodant la fin de la ligne de l'édition ou du manuscrit saisi. Pour ne pas perdre cette information, la balise <s> a été insérée à la fin de chaque ligne, ce qui est satisfaisant dans le cas des textes en vers (<s> délimite donc les vers), mais produit un découpage philologique et linguistiquement arbitraire des 54 textes en prose. L'unité sémantique et structurale intéressant la plupart des utilisateurs d'un corpus étant la phrase, nous avons introduit manuellement la ponctuation dans les 32 textes en prose tirés d'une édition imprimée disponible (ces textes sont donc classifiés par les attributs ‹*vers=«non»› et ‹ponctuation=«oui»›).

Au niveau lexical, la balise <word> délimite la forme graphique et contient les attributs suivants :

- la valeur de « deespos » est l'étiquette numérique des fichiers originaux (p. ex. « 566 », cf. tableau 10) ;
- la valeur de « pos » est la traduction de l'étiquette numérique dans « deespos », p. ex. « VER:futu:3:pl » ;
- la valeur de « taggerpos » est la catégorie sélectionnée automatiquement par l'étiqueteur (cf. section 4.2). Pour améliorer la lisibilité, le tableau 10 ne montre pas cet attribut, qui peut être intéressant pour comparer les résultats d'une analyse manuelle à ceux d'une analyse automatique ;
- la valeur de « lemma » est le lemme (ou une série de lemmes dans les cas ambigus que l'étiqueteur ne peut pas résoudre – p. ex. si la partie du discours est identique) ;
- la valeur de « src » indique l'origine du lemme, c'est-à-dire la ressource lexicale d'où il provient (cf. section 4.2).
- De manière non systématique nous avons utilisé l'attribut « note » pour indiquer les modifications apportées aux données originales. Ainsi, les quelque 54.000 signes de ponctuation insérés manuellement sont systématiquement marqués par ‹note=« ajout »›, et certaines fautes de frappe dans l'annotation de Dees ont été corrigées et marquées par ‹note=« tag 018 corrected »› ou ‹note=« deespos inconnu »›, etc.

Certes, l'annotation XML de la version actuelle pourrait être améliorée, mais les ressources dans le cadre de ce projet étant limitées, pour cette première version du corpus nous avons accordé la priorité à la lemmatisation. D'autre part les versions futures du Nouveau Corpus d'Amsterdam bénéficieront sans doute des efforts de documentation et de standardisation qui seront poursuivis dans le cadre du Consortium international pour les corpus de Français Médiéval (CCFM), qui réunit les détenteurs de corpus de français ancien et la majorité des collaborateurs à ce volume.

BIBLIOGRAPHIE

Blumenthal, Peter & Stein, Achim (éd.) (2002) : Tobler-Lommatzsch : Altfranzösisches Wörterbuch. 4 CD-ROMs und DVD mit Begleitbuch, Stuttgart : Steiner.

Carolus-Barré. L. (1952) Compte rendu de Gossen, Charles Théodore (1949), Petite Grammaire de l'ancien picard. Romania 73, 109–118, 511–513.

Carolus-Barré, L. (1964) Les plus anciennes chartes en langue française, 1.1, Recueil des pièces originales conservées aux Archives de l'Oise 1241–1286, Paris.

Dees, Anthonij (1971) : Etude sur l'évolution des démonstratifs en ancien et en moyen français, Groningen : Wolters-Noordhoff.

Dees, Anthonij, avec le concours de Pieter Th. Van Reenen et de Johan A. De Vries (1980) : Atlas des formes et des constructions des chartes françaises du 13e siècle, Tübingen : Niemeyer.

Dees, Anthonij, avec le concours de Marcel Dekker, Onno Huber et Karin Van Reenen-Stein (1987) : Atlas des formes linguistiques des textes littéraires de l'ancien français, Tübingen : Niemeyer.

Heiden, Serge & Prévost, Sophie (2005): « Étiquetage d'un corpus hétérogène de français médiéval: enjeux et modalités » – Kabatek, Johannes & Pusch, Claus & Raible, Wolfgang (ed.): Romance Corpus Linguistics II: Corpora and Diachronic Linguistics, Tübingen: Narr.

Labov, William (1982) : Building on empirical foundations 1982.

Schmid, Helmut (1994) : « Probabilistic Part-of-Speech Tagging using Decision Trees » – Sima'an, K. & Bod, R. & Krauwer, S. & Scha, R. (éd.) : Proceedings of the International Conference on New Methods in Language Processing (NeMLaP'94), Manchester September 1994, Manchester : UMIST.

Schwan-Behrens (1913) : Grammaire de l'ancien français, troisième partie : matériaux pour servir d'introduction à l'étude des dialectes de l'ancien frlançais publiés parDietrich Behrens, Leipzig : Reisland.

Stein, Achim (2003) : « Étiquetage morphologique et lemmatisation de textes d'ancien français » – Kunstmann, Pierre et. al. (éd.) : Ancien et moyen français sur le Web : Enjeux méthodologiques et analyse du discours, Ottawa : Les Éditions David, 273–284.

Stein, Achim & Schmid, Helmut (1995) : « Etiquetage morphologique de textes français avec un arbre de décisions » – traitement automatique des langues, Volume 36, Numéro 1–2 : Traitements probabilistes et corpus, 23–35.

Van Reenen, Pieter (1976) : Taalgeografisch onderzoek naar het Frans in de Middeleeuwen, een kwantitatieve benadering, School voor Taal- en Letterkunde, 's-Gravenhage.

Van Reenen, Pieter (1989): « La pertinence linguistique des rimes en EN/AN dans la Bible de Macé de la Charité » – Actes du Colloque International sur l'Ancien Provençal, l'Ancien Français et l'Ancien Ligurien (Nice, septembre 1986): Bulletin du Centre de Romanistique et de Latinité Tardive, no double 4–5, janvier 1989, 247–266.

Van Reenen, Pieter, avec le concours de Margôt van Mulken et Evert Wattel (2007): Chartes de Champagne en français conservées aux Archives de l'Aube 1270–1300, Orléans: Paradigme. I–XV, 1–283.

Van Reenen, Pieter & Schøsler, Lene (2000) : « Corpus et stemma en ancien et en moyen français. Bilan, résultats et perpectives des recherches à l'Université libre Amsterdam et dans les institutions collaboratrices » – Buridant, Claude (éd.) : Le moyen français. Le traitement du texte, Strasbourg : Presses universitaires de Strasbourg, 25–54.

ADRESSES INTERNET

Dictionnaire étymologique de l'ancien français (DEAF) :
http://www.deaf-page.de
Laboratoire de Français Ancien (LFA, Ottawa) :
http://www.uottawa.ca/academic/arts/lfa/
Nouveau Corpus d'Amsterdam (site de téléchargement) :
http://www.uni-stuttgart.de/lingrom/stein/corpus
Tobler/Lommatzsch *Altfranzösisches Wörterbuch*, version informatisée :
http://www.uni-stuttgart.de/lingrom/stein/tl/
TreeTagger : paramètres pour l'ancien français
http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html