Modèles théoriques inductifs et propositions d'applications aux donnés textuelles de l'ancien français

Xavier-Laurent Salvador [1], Fabrice Issac [2]

[1] Université de Paris XIII, Laboratoire LDI (Lexique Dictionnaire et Informatique) – UMR CNRS 7187, 99 av. Jean-Baptiste Clément, 93430- Villetaneuse

[2] Université de Paris XIII, Laboratoire LDI (Lexique Dictionnaire et Informatique) – UMR CNRS 7187, 99 av. Jean-Baptiste Clément, 93430- Villetaneuse

Résumé

Le présent article définit la pertinence d'une approche inductive et cumulative du traitement automatique de l'ancienne langue française. Les auteurs exposent ensuite les applications pratiques réalisées ou, le projet étant en cours de développement, envisageables à partir de l'outil développé.

Abstract

The present paper aims at demonstrating the adequacy of an inductive and cumulative approach of the automatic processing of medieval French texts as at LDI laboratory. First the authors show what has been done and, for this project is still under development, what could be done with their tool.

Key-words: medieval French, induction, automatic labeling

1. Introduction

La méthode de constitution d'une ressource lexicale que nous développons dans le présent article est fondée sur une approche inductive et cumulative. Nous montrons dans un premier temps l'exploitation de la force des approches combinatoires de l'outil informatique basé sur les modèles fournis par les théories de la grammaire de l'ancien français pour prédire la somme envisageable de toutes les réalisations des formes extraites des répertoires de l'ancien français. Nous montrons ensuite les outils développés par nos soins afin de décrire la morphologie des parties du discours. Enfin, nous présentons quelques réalisations sous forme d'autoévaluation appliquées au champ théorique de la conjugaison.

2. Pertinence des objectifs au regard des ressources existantes

2.1. Les travaux existants

Plusieurs équipes, en France et à l'étranger, proposent des ressources électroniques pour l'ancienne langue. Ces travaux ont pour objectif de proposer des *corpus* permettant d'étudier la langue à l'aide des outils de traitement automatique des langues. La tâche de construction de *corpus* consiste tout d'abord à réunir et à structurer de manière uniforme des textes selon un

schéma cohérent. L'ancienne langue pose, de ce point de vue, un certain nombre de problèmes spécifiques liés au travail particulier appliqué par les éditeurs dans le cadre des éditions critiques ou diplomatiques, et selon le souci de l'auteur d'intégrer les informations dialectales à son travail. La Base du Français Médiéval ou BFM, par exemple, propose ainsi des éditions critiques de textes structurés, en utilisant les recommandations TEI. Ce corpus est utilisable au travers de l'outil Weblex (Heiden and Lavrentiev, 2004). Dans le schéma constitution/ édition de corpus puis de leur utilisation au travers d'outils de concordance ou de mesures lexicométriques, nous notons qu'il manque l'étape d'annotation. Celle-ci consiste à enrichir le texte édité d'informations morphologiques, syntaxiques, sémantiques, topochronologiques. L'enrichissement linguistique nécessite l'utilisation de ressources propres qui, si elles existent pour le français moderne, semblent faire défaut pour l'ancien français. Pour pallier ce manque, plusieurs stratégies sont utilisées. La première consiste à utiliser les lexiques et dictionnaires disponibles (essentiellement le Tobler-Lommatzsch et le Godefroy) ainsi que différents corpus déjà étiquetés pour constituer une ressource dictionnairique lemme/forme fléchie/étiquette. Un important travail de normalisation est nécessaire pour harmoniser, et rendre utilisable, la ressource. Stein (2003) utilise une telle ressource pour procéder à un étiquetage via le *Treetagger*. Prévost et Heiden (2002) utilisent SATO, un outil permettant entre autre de construire des règles d'étiquetage basées sur l'analyse des désinences. A la suite de ce premier travail, des méthodes inspirées de la technique d'E. Brill ont été utilisées pour compléter l'étiquetage. Dans l'ensemble, nous constatons que les méthodes utilisées se focalisent sur la réutilisation des ressources disponibles et comblent les manques via l'utilisation d'analyseurs de surface essentiellement morphologiques et de techniques statistiques.

3. La démarche inductive appliquée à l'exemple de la conjugaison du présent de l'indicatif des verbes de la classe I [1]

3.1. La variation des bases à l'intérieur du paradigme entraîne une difficulté de prédiction des formes rencontrées

La variation morphologique des bases au sein des conjugaisons de l'ancien français illustre l'un des problèmes spécifiques liés au traitement de l'ancienne langue. Nous voudrions illustrer la problématique de la démarche inductive appliquée à la conjugaison des verbes de classe I afin de montrer l'originalité des solutions que nous envisageons. Au contraire de conjugaisons authentiquement françaises, comme le futur ou le conditionnel qui ne connaissent pas d'alternances accentuelles – ce sont des temps faibles – le temps du présent de l'indicatif permet de souligner une difficulté fondamentale dans l'optique du traitement automatique de la langue, à savoir la résolution du lien morphologique implicite – perçu comme tel par le sujet parlant – unissant les formes actualisées d'un paradigme protéiforme. Le problème se pose donc de la manière suivante : en ancien français, les formes « *aim* » et « *amons* » appartiennent au même paradigme de la conjugaison d'un verbe dont on s'accorde à dire que l'infinitif canonique est « *amer* », mais qui pourrait tout aussi bien être actualisé en discours sous une forme analogique construite sur le modèle de P1, à savoir la forme du français moderne « *aimer* ».

[1] Nous rappelons que selon la terminologie ordinairement usitée en philologie romane, la classe I correspond aux verbes issus de la famille des verbes latins en *-are.* Nous avons par ailleurs conscience qu'il existe depuis une vingtaine d'années des outils de conjugaison opérants. Toutefois, l'application de la méthode au champ de la conjugaison ne résume pas l'ambition de notre ressource. Elle en constitue un premier champ d'application dont il est facile de contrôler le développement par le recours aux sources externes.

3.2. La démarche inductive est pertinente en contexte

Le comportement précédemment décrit trouve largement son explication dans le commentaire historique de la morphologie du français. Rappelons en effet que le maintien en général [2] de l'accent tonique dans les formes verbales du latin au français moderne a entraîné des altérations morphologiques des conjugaisons qui expliquent les nombreuses variations qui affectent le comportement de la base verbale. Nous n'avons donc d'autre alternative si nous voulons concevoir un outil efficace en phase avec la théorie que de calculer l'ensemble du lexique au sein d'une ressource produisant une langue fantôme contenant tous les possibles. C'est un postulat méthodologique qui implique : (i) le choix d'une description étendue comprenant l'ensemble des formes de l'ancienne langue (ii) l'inclusion exhaustive des produits: il n'y a pas, comme pour le français moderne, une notion de forme correcte opposée à une forme incorrecte. De fait, toute forme est correcte si elle existe dans le corpus fini des textes, et est incorrecte si elle n'est pas attestée, (iii) le respect des produits surnuméraires. Toute forme est considérée en attente d'une éventuelle rencontre avec son attestation jusqu'à ce que le parcours de toute la production de l'ancienne langue puisse être accompli, (iv) le perpétuel amendement de la ressource.

4. La description de la nomenclature basée sur le plus petit dénominateur morphologique commun

4.1. Constitution de la ressource

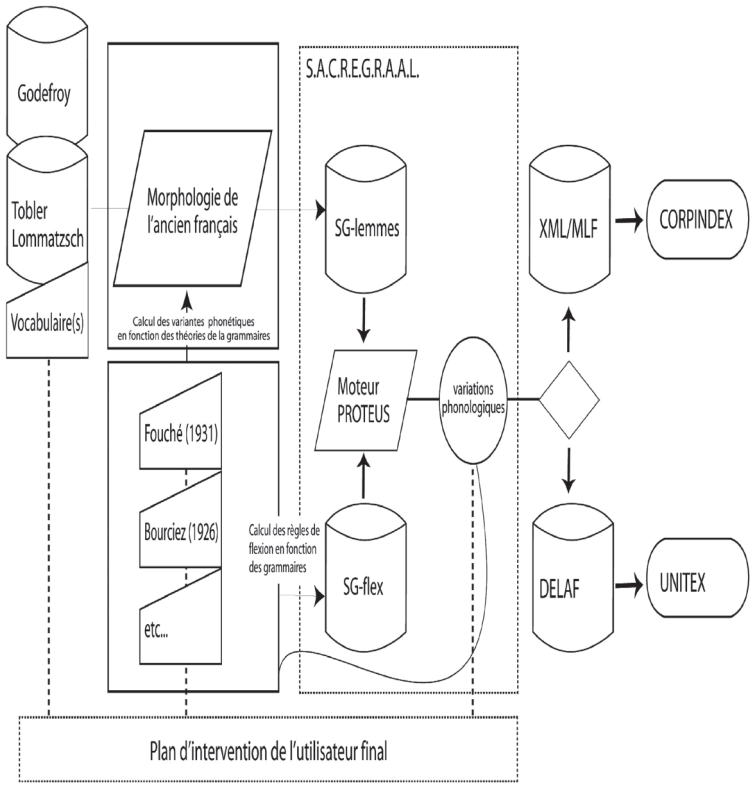
Pour la constitution de la ressource dictionnairique, nous sommes partis de la nomenclature de deux répertoires de l'ancien français, les dictionnaires de l'ancien français de Tobler-lommatzsch d'une part et de Godefroy de l'autre réunis, de sorte que le dictionnaire vient combler les lacunes du lexique. Plutôt que de procéder à l'unification de la ressource en créant une liste unie où les termes absents de l'un seraient remplacés par l'extraction du complément dans le second, nous avons choisi de modifier la nomenclature de notre propre ressource afin que tous les mots du lexique et du dictionnaire soient présents en même temps dans la base lemmatisée. Ce procédé, qui alourdit considérablement la charge du calculateur, présente trois avantages: dans un premier temps, cela permet de confronter l'originalité de deux nomenclatures de manière à constituer une base originale complète ; dans un second temps, cela permet également d'enregistrer de manière non artificielle l'ensemble des variantes d'un même lemme telles qu'elles figurent de manière originale dans l'un des deux dictionnaires et pas dans l'autre sous la même entrée. Enfin, il nous a semblé pertinent de pouvoir mesurer la fréquence d'enregistrement d'un terme dans les répertoires [3]. La ressource est constituée en trois étapes. (i) Dans un premier temps, on isole de la base de données l'ensemble des verbes se terminant en « -er ». On ajoute à chacune d'entre elle une étiquette de description morphologique « vb.g1 », et marque l'ensemble des verbes type « laisser » qui prennent étymologiquement un « e » à la première personne du singulier. (ii) Après avoir procédé au recensement des verbes et de leurs variantes, il s'agit d'identifier dans chacune des formes la voyelle tonique autour de laquelle se produira l'accident phonétique le cas échéant. Chaque syllabe tonique est marquée par un méta-caractère dont nous avons convenu du sens. (iii) Dans un troisième temps, toutes les variations vocaliques associées

[2] Rappelons toutefois que de nombreux changements d'accentuation ont eu lieu dans l'évolution de la langue. Sur ce sujet, voir notamment Fouché P. (1976 : 3-12).

[3] Dans la perspective où se posera vite le problème de la résolution de l'ambiguïté polysémique de certains termes, il n'est certainement pas anodin de prendre en considération la présence d'une unité dans les deux dictionnaires ou dans un seul des deux, et de savoir lequel.

à la voyelle tonique sont envisagées *en fonction d'une liste tabulaire extérieure au programme*. De plus, selon qu'elle est libre ou entravée, et selon la nature de la consonne précédente et subséquente, il est possible de prédire un ensemble d'accidents altérant la longueur de la base, aussi bien les réductions que les allongements. (iv) Enfin, dans un quatrième et dernier temps, nous produisons pour chacune des formes ainsi dérivées du lemme d'origine un ensemble de variantes phonologiques indexées sur une seconde liste tabulaire extérieure au programme d'origine nous permettant d'envisager toutes les variantes de graphies des sons nasalisés, des consonnes vélaires ou bien entre autre les équivalences de graphies à vocation morphologique équivalentes « s/z », « x/us/ls » pour ne citer que les plus évidentes.

Le schéma 1 illustre les différentes étapes du projet. En entrée, le lexique d'origine alimente le moteur de flexion qui calculent l'ensemble des données en fonction des paramètres saisis par l'utilisateur final. La ressource produite peut être projetée sur un corpus. Ce protocole d'ores et déjà en application offre une grande souplesse d'emploi et permet à chaque instant du repérage d'un corpus de pouvoir modifier les indices fondamentaux qui conditionnent le résultat obtenu. Si l'on nous permet à nouveau d'illustrer notre propos, nous dirions que la machine laisse couler toute la puissance de la langue pendant que nos indices tabulaires extérieurs en contrôlent le flux d'écoulement.



Schema 1 : Etapes du projet

4.2. La valeur hyperlemme

De manière générale, l'hyperlemme est une valeur attachée à un ensemble de formes, puisque il ne peut pas exister en dehors de ces formes. Nous le considérons comme le plus petit dénominateur commun morphologique. Nous voyons en lui l'informatisation d'un donné empirique: l'intuition de la langue. La valeur hyperlemme est attachée aux hypostases de la forme en langue telles qu'on les rencontre dans les lexiques. Notre base de données, qui se présente donc sous la forme d'un

répertoire de lemmes (« SG-lemmes ») associés à une valeur-clé, se compose à la base de la réunion des deux lexiques de l'ancien-français: le Tobler-Lommatzsch, *Altfranzösisches Wörterbuch* et le *Dictionnaire de l'ancien français* de Godefroy. La nomenclature de l'ouvrage de Frédéric Godefroy qui ouvre plus de 160000 entrées empruntées à tous les dialectes et tous les vocabulaires de l'ancienne langue offre, à l'image du Dictionnaire de Du Cange pour le latin médiéval, une base puissante pour la génération de la langue standard. Le lexique de Tobler et Lommatzsch intègre plus de 30 000 variantes renvoyant à d'autres entrées. SG-lemme enfin, ce sont en quelques chiffres et à l'heure où nous nous employons à en illustrer les perspectives d'application, source TL : 76908 entrées; source GD : 55932; nombre de lemmes : 132.866 et nombre d'hyperlemmes : 82.811.

5. Le code Proteus utilisé pour la constitution d'une ressource lexicale exhaustive

5.1. Les solutions existantes

Parmi les outils existant permettant de représenter et d'utiliser une ressource lexicale nous retiendrons, parce qu'ils proposent la plateforme la plus complète, Unitex/Intex et Nooj.

Leurs dictionnaires pour les formes simples se décomposent en deux entités appelées DELAS et DELAF. Le premier contient la forme canonique des mots ainsi qu'un code permettant de réaliser la flexion. Le DELAF est l'ensemble des formes fléchies généré à partir du DELAS et des règles de flexion. Une entrée du DELAS se décompose de la manière suivante :

<lemme>, <code de transformation>+<traits>

Le code de transformation fait référence à une grammaire qui est en fait un ensemble d'opérations sur des caractères. Celles-ci sont au nombre de 4 et permettent de déplacer une lettre à partir de la fin d'un mot (L), de remettre une lettre déplacé (R), de dupliquer une lettre (C), de supprimer une lettre (D). En plus de ces opérations il est possible d'ajouter une ou plusieurs lettres. Le code transformation LLLDèRRRais permet de transformer *céder* en *cèderais*.

Nooj est le successeur d'Intex. Parmi les nouveautés qu'il intègre, notons une modification complète de l'architecture des dictionnaires [4]. Le nouveau modèle proposé permet de gérer d'une seule manière les mots simples et les mots composés, toujours à l'aide des opérateurs de manipulation des caractères auxquels ont été ajoutés des opérateurs de manipulation de mots. Ceux ci permettent d'*aller à la fin du mot précédent* ou d'*aller à la fin du mot suivant*. La méthode utilisée mélange manipulation de caractères et utilisation de *grammaires morphologiques*.

Ces outils, très puissants par ailleurs, nous sont apparus lacunaires au regard de la spécificité du traitement de l'ancienne langue pour les raisons suivantes: (i) Une description *in extenso* de la langue ne permet pas d'effectuer l'analyse des mots nouveaux mais ceux dont la flexion est régulière. Il faut donc mettre en place un modèle de flexion qu'il sera possible d'utiliser aussi bien en analyse qu'en génération. Il sera ainsi possible de faire de l'analyse non seulement sur les mots connus mais aussi sur les mots inconnues. (ii) Une règle de flexion, appliquée sur un mot, n'est jamais complètement autonome, mais s'inscrit dans le cadre d'un ensemble. On regroupera par exemple dans un même groupe l'ensemble des conjugaisons d'un type de verbe, et ce pour tous les temps. Cette manière de faire va faciliter la conception, la maintenance et l'utilisation des règles. (iii) L'ancien français présente de nombreuses variétés dialectales. Au sein d'un même ensemble de conjugaison, nous devons définir une série de sous ensembles qui décrivent la réalité d'un état de langue.

[4] Voir Silberztein (2005).

5.2. Le modèle Proteus

5.2.1. Description des opérations de flexion au sein de Proteus

Nous avons donc décidé de développer une application dédiée baptisée Proteus. Le modèle proposé devait s'appuyer sur un ensemble d'outils simples, devait également pouvoir s'intégrer facilement au sein d'applications tierces et permettre enfon une utilisation des dictionnaires produits dans un au autre environnement. Ce dernier s'appuie (i) sur un ensemble d'opérateurs sur les caractères et (ii) d'un objet que nous appelons « pile » capable de recevoir des caractères que l'on écarte et dont on se souvient. L'utilisation dans un certain ordre de ces opérateurs constitue un « code » correspondant à une fonction de transformation : mot + code = mot fléchi. En simplifiant le modèle à l'extrême, la création d'une règle permettant de générer une forme fléchie à partir d'un lemme se fait (i) en mettant des lettres de côté ou en réinsérant des lettres dans un mot (ii) en effaçant ou en insérant des lettres. Par défaut, les opérations s'appliquent sur les caractères placés en fin de mot ou, selon l'opérateur, en haut de la pile de caractères mis de côté. Les opérations décrivent les transformations à effectuer pour obtenir une forme fléchie à partir d'un lemme, c'est-à-dire pour construire un lexique de formes fléchies. Pour une utilisation en analyse, le sens de chaque opérateur est inversé. Les différents opérateurs exprimant une transformation sont les suivants :

- P (emPile) : déplace un caractère de la liste vers la pile
- D (Dépile) : déplace un caractère de la pile vers la liste
- E (Efface) : efface un caractère de la liste.

Ce code permet donc la flexion des verbes du type de *céder*. Cependant, il n'est pas possible avec ce type de code d'effectuer l'opération inverse, *i.e.* retrouver le lemme à partir de la forme fléchie, du fait de la présence de l'opérateur E (efface) qui, au contraire des autres opérateurs, n'est pas réversible. C'est pourquoi nous ajoutons un opérateur d'effacement prenant en paramètre le ou les caractères à effacer.

- \zyx\ (Efface) : efface des caractères de la liste.

Le code de flexion de notre précédent exemple devient donc PPP\é\/è/DDD/ais/. Pour simplifier l'écriture du code, il est possible d'indiquer le nombre de répétitions d'un opérateur avant celui-là. Le code PPP\é\/è/DDD/ais/ peut donc s'écrire 3P\é\/è/3D/ais/.

La gestion des préfixes/suffixes nécessite aussi l'ajout d'opérateurs :

- R (Remplit) : transfère tous les caractères de la liste vers la pile
- V (Vide) : transfère tous les caractères de la pile vers la liste.

L'opérateur R permet de préparer un ajout en début de mot puisque tous les caractères sont « mis de côté » dans la pile. Nous sommes maintenant en mesure de décrire les flexions de la forme : « planter > sousplantons ». La transformation « enlever les caractères 'er' en fin de mot, ajouter 'sous' en début de mot puis 'ons' en fin de mot » s 'écrit donc \re\R/sous/V/ons/.

- P|x| (emPile) : déplace les caractères de la liste vers la pile jusqu'à rencontrer le caractère x.

La modification de l'opérateur d'empilement permet ainsi d'accéder directement aux mots simples de l'expression. A noter que l'accès à différents éléments d'une expression se fait par empilement et dépilement successifs. Un dernier opérateur est ajouté pour permettre d'effacer l'ensemble des caractères de la pile. Cet opérateur a été ajouté essentiellement pour exprimer le fait qu'une flexion n'existe pas. Ainsi, le code 'RS', qui remplit la pile puis l'efface, permet cette action. Nous verrons dans quels cas nous l'utiliserons.

5.2.2. Regroupement des expressions de codes de flexion au sein de Proteus

Cependant il est bien souvent nécessaire de regrouper plusieurs transformations, masculin/féminin – singulier/pluriel pour les noms et les adjectifs, l'ensemble des temps pour les verbes. Il doit être possible d'effectuer les regroupements standards, mais aussi de gérer les variantes, comme pour l'ancien français, et de focaliser sur une représentation partant d'une description générale pour ensuite décrire les exceptions. En ce sens, nous nous rapprochons d'une description « à la Bescherelle » qui donne en premier lieu la conjugaison prototypique, puis énumère les exceptions.

Prenons comme exemple la conjugaison d'un verbe du premier groupe à l'imparfait. La terminaison prototypique peut être donné de la manière suivante pour l'ancien français:

```
<flex id="v1ii" type="term">
        <name>Vmii</name>
        <info>verbes indicatif imparfait base XI</info>
                <flex id="p1ns">
                        it<name>1s-</name>
                        <code>/eie/</code>
                </flex>
                <flex id="p2ns">
                        <name>2s-</name>
                        <code>/eies/</code>
                </flex>
                <flex id="p3ns">
                        <name>3s-</name>
                        <code>/eit/</code>
                </flex>
                <flex id="p1np">
                        <name>1p-</name>
                        <code>/iiens/</code>
                </flex>
                <flex id="p2np">
                        <name>2p-</name>
                        <code>/iiez/</code>
                </flex>
                <flex id="p3np">
                        <name>3p-</name>
                        <code>/eient/</code>
                </flex>
</flex>
```

Chaque flexion a un nom (name), un code au sens Proteus (code) et un identifiant (id). La précédente définition stipule qu'au présent de l'indicatif pour les verbes du premier groupe, on ajoute en fin de mot *eie*, *eies*, *eit*, *iiens*, *iiez*, *eient*. Il en est de même pour les autres temps puis les autres modes. Dans un deuxième temps nous effectuons des regroupements pour former des classes :

```
<flex id="vig1-1" type"nonterm">
        <name></name>
        <info>premier groupe verbe type aimer indicatif</info>
                <op type="add">
                        <item value="v1ip"/>
                        <item value="v1ii"/>
                        <item value="v1ips"/>
                        <item value="v1ifs"/>
                </op>
</flex>
```

À cet ensemble de définitions, nous pouvons prendre en compte les variations dialectales et diachroniques en appliquant un masque qui modifie l'ensemble des codes de la classe. Cette modification se fait éventuellement de manière sélective sur les codes de flexion. La transformation du code Proteus se fait en appliquant un code Proteus [5] de la sorte :

```
<mask id="modifXIII">
        <info>modification XIII imparfait</info>
        <item ervalue="\.p1ns">P3E/eies/D</item>
        <item ervalue="\.p2ns">2PEPE/a/V</item>
        <item ervalue="\.p3n[sp]">R2DE/a/V</item>
        <item ervalue="\.p1np">R2DEV</item>
</mask>
```

Les codes à modifier sont identifiés par une expression rationnelle (attribut 'ervalue') sur l'identifiant de la flexion : '\.p1ns' s'applique à toutes les flexions des premières personnes du singulier. En l'occurrence, l'imparfait se prête bien à la démonstration : c'est un temps faible (toujours accentué sur la désinence) caractérisé par une voyelle thématique dont le timbre ou la nature ont changé selon les époques ou les lieux. Le masque d'identifiant 'modifXIII' génère le code qui procède à l'infixation de la voyelle thématique correspondant à la conjugaison du XIIIe siècle de l'imparfait et à l'ajout des désinences de personnes. Il est susceptible d'être sérialisé et de produire en synopse tous les paradigmes du sous-ensemble des variantes de l'imparfait.

Le code suivant illustre la génération de la conjugaison de l'imparfait du XIIIe par l'application des masques successifs 'modifXI' (modification de la conjugaison standard caractéristique du XIe siècle), 'modifXII', 'modifXIII', 'vrber' (détermination de la base).

```
<flex id="vg1i-4" type="final">
    <name></name>
    <info>premier groupe imparfait XIIIe a-1</info>
        <op type="mask" value="v1ii">
                <item value="modifXI"/>
                <item value="modifXII"/>
                <item value="modifXIII"/>
                <item value="vrber"/>
        </op>
</flex>
```

5.3. Le modèle inductif appliqué à la génération massive

L'analyse lexicale consiste, dans le cas d'un document électronique, à (i) regrouper des caractères en unités lexicales, (ii) attacher à ce regroupement un ensemble d'informations (lemme, partie du discours, morphologie). La réalisation de cet étiquetage se fait via l'utilisation de deux ressources. Tout d'abord, notre dictionnaire de lemmes où chaque hyperlemme est associé à un lemme et à l'étiquette décrivant la partie du discours à laquelle il appartient et sa source. Il se présente par exemple de la façon suivante :

amer	amer	Vrb.	TL
aimer	amer	Vrb.	TL:G:P

[5] Cette mise en abîme semble inadaptée, puisque un code Proteus a été conçu pour s'appliquer sur un élément de langue. Cependant il nous a paru inadéquat d'introduire une nouvelle syntaxe de transformation. Une conséquence de ce choix, c'est la présence de caractères d'échappement rendant certaines règles de transformations humainement délicates à interpréter.

Le verbe « amer » est présent dans le Tobler-Lommatzsch (TL) et se rattache naturellement à l'hyperlemme « amer ». Après traitement phonétique de la base SG-lemmes symbolisé « :P », nous avons généré par variantes après isolement de la tonique la forme « aimer » rattachée à l'hyperlemme « amer » de TL. Le même traitement est appliqué à la forme attestée dans Godefroy (« :G ») et les deux étiquettes sont regroupées (« :TL:G:P »).

Ensuite, un dictionnaire de flexions au format Proteus. Dans le cas présent par exemple, il est important que nous récupérions après traitement par le code Proteus de la base lexicale les formes « amons », « aimons », « ame » et « aime ». Les lemmes ainsi produits sont rattachés à l'hyperlemme source:

aimions amer Vrb.(P4 imparfait indicatif) TL:G:P

Il est intéressant de noter que ces deux ressources pourraient être utilisées, c'est-à-dire projetées sur un texte, de deux manières différentes, soit en analyse (i) soit après génération (ii). L'analyse (i) consiste à appliquer pour un mot du texte les symétriques des différentes règles de flexion pour obtenir un lemme potentiel. Il ne reste plus ensuite qu'à vérifier dans le dictionnaire de lemmes si celui-là existe réellement, ou non. La seconde méthode (ii) consiste à générer *a priori* l'ensemble des formes fléchies. On procède ensuite à l'ajout d'une étiquette en comparant directement l'élément lexical avec la liste de mots fantômes ainsi produite. Notre choix se porte sur la génération inductive *a priori* de sorte que dans l'espace imposé par un lexique clos, la gestion des flexions et des variantes est incluse par définition dans la théorie. Nous adoptons une démarche modulaire de filtres associés qui permet d'envisager que chacun puisse au gré de ses besoins et des évolutions de la science de la philologie adapter la ressource par l'adaptation des règles. Ainsi, si nous ne prétendons pas tout prévoir ni tout connaître de tous les états de langue ni de tous les dialectes, nous proposons un outil flexible basé sur une langue idéale – celle des grammaires de l'ancien français – susceptible de s'adapter aux besoins, des plus précis : le calcul de toutes les formes de l'ancien bourguignon dans une région précise à une époque donnée en fonction des règles définies par le chercheur concerné ; au plus ambitieux : le calcul de toutes les formes du picard du IXe au XIIIe siècle. Le développement de lexiques pour Unitex rend également possible non seulement la modélisation des états de langue, mais leur schématisation incluse dans une chronolocalisation. La constitution d'une ressource dictionnairique telle que nous l'envisageons sur le modèle inductif que nous décrivions en début de ce chapitre serait donc à la fois un outil susceptible de produire des résultats, mais proposerait également un protocole théorique d'approche des états des anciennes langues. L'objectif que nous nous sommes imposé est de fournir une ressource utilisable pour différents outils, au premier rang desquels un dictionnaire de lemme au format XML (respectant le standard LMF [6]). Afin d'illustrer ce propos, nous avons produit une version DELAF pour que le résultat soit exploitable sous Unitex.

Références

Andrieux N. and Baumgarner E. (1983). *Système morphologique de l'ancien français. A. Le verbe*. Paris : Klincksieck.

Bourciez E. (1926). *Précis historique de phonétique française,* 6e éd. Paris.

Brunot F. (1886). *Grammaire historique de la langue française*. Paris.

[6] *Lexical Markup Framework* est un modèle de structuration de données linguistiques, notamment lexicales, dont l'objectif est de favoriser la réutilisation de ces ressources (voir http://www.lexicalmarkupframework.org/).

Chabaneau C. (1878). *Histoire et théorie de la conjugaison française*. Paris.

Chambon J.-P. (1997). Les emprunts du français moderne aux dialectes et aux patois: une illusion d'optique en lexicologie française historique. In *Lalies. Actes des sessions de linguistique et de littérature*, 17, pp. 33-53.

Fouché P.(1931). *Le Verbe français*. Paris.

Fouché P. (1976). *Morphologie historique du français – Le Verbe*. Paris : Klincksieck.

Heiden S. and Lavrentiev A. (2004). Ressources électroniques pour l'étude des textes médiévaux : approches et outils. *Revue française de linguistique appliquée*, IX(1) : 99-118.

Meiller A. (1991). Pour compléter et corriger le Tobler-Lommatzsch. *Romania*, 112 (3,4) : 533-540.

Meyer-Lübke W. (1908). *Historische Grammatik der franzoesischen Sprache*. Heidelberg.

Pope M.K. (1952). *From Latin to Modern French.* Manchester : Manchester University Press.

Prévost S. and Heiden S. (2002). Etiquetage d'un corpus hétérogène de français médiéval : enjeux et modalités. In Pusch, C.D. and Raible, W., editors, *Romanistische Korpuslingustik: Korpora und gesprochene Sprache, Romance Corpus Linguistics: Corpora and Spoken Language*, 1st Freiburg Workshop on Romance Corpus Linguistics, Freiburg, 6 - 8 Octobre 2000, Tübingen : Gunter Narr Verlag, pp. 127-136.

Rohlfs G. (1949). *Historische Grammatik der italienischen Sprache und ihrer Mundarten*. Bern : Francke.

Silberztein, M. (2005). *NooJ's Dictionaries.* In *the Proceedings of LTC 2005*, Poznan.

Skarup P. (1994). *Morphologie Synchronique de l'ancien français, Etudes romanes*. Copenhague : Museum Tusulanum.

Stein A. (2003). Étiquetage morphologique et lemmatisation de textes d'ancien français. In Kunstmann, P., Martineau, F. and Forget, D., editors, *Ancien et moyen français sur le Web: Enjeux méthodologiques et analyse du discours*, Ottawa: Les Éditions David, pp. 273-284.

Straka G. (editor) (1972). *Les Dialectes de France du Moyen Âge et aujourd'hui : domaine d'oil et domaine franco-provençal*. Paris : Klincksieck.

Walker Douglas C. (1981). Old French Morphophonology. *Studia Phonetica* 19, Ottawa: Didier.