LGeRM Lemmatisation des mots en Moyen Français

Gilles Souvay, Jean-Marie Pierrel

LGeRM

Lemmatisation des mots en moyen français

Gilles Souvay — Jean-Marie Pierrel

Analyse et Traitement Informatique de la Langue Française
CNRS & Nancy-Université
44, avenue de la Libération
B.P. 30687
F 54063 NANCY CEDEX
gilles.souvay@atilf.fr
jean-marie.pierrel@atilf.fr

RÉSUMÉ. *Contrairement à la plupart des langues modernes, le moyen français est une langue dont l'orthographe n'est pas encore stabilisée. Il existe de très nombreuses variantes pour un même mot et en conséquence les méthodes classiques de lemmatisation ne peuvent pas s'appliquer. LGeRM (Lemmes, Graphies et Règles Morphologiques) propose une solution qui s'appuie sur une base de formes connues lemmatisées et sur un ensemble de règles graphémiques et morphologiques spécifiques de la langue médiévale. Il permet ainsi de faciliter la consultation d'un dictionnaire, l'interrogation et la lemmatisation de textes médiévaux et trouve des applications dans l'édition électronique de manuscrits et la construction automatique de glossaires. Cet outil polyvalent est accessible sur internet à l'adresse www.atilf.fr/dmf.*

ABSTRACT. *Unlike most modern languages, Middle French is a language whose spelling is not yet stabilized. There is a great deal of variation in the spelling of a word and accordingly the traditional methods for lemmatization cannot be used. LGeRM (Lemmes, Graphies et Règles Morphologiques) proposes a solution based on a databank containing known lemmatized spellings and a set of graphical and morphological rules specific to the medieval language. LGeRM can provide help in consulting a dictionary, browsing or lemmatizing medieval texts, and it can be useful in the electronic edition of manuscripts and the automatic construction of glossaries. This multipurpose tool is accessible on the Internet at www.atilf.fr/dmf.*

MOTS-CLÉS : *Lemmatisation, dictionnaire, glossaire, moyen français, ancien français*

KEYWORDS: *Lemmatization, dictionary, glossary, middle French, old French.*

1. Introduction

1.1 *Contexte de la recherche*

L'utilisation de corpus informatisés, textes ou dictionnaires, est devenue capitale pour les études linguistiques. Pour détecter ou expliquer un phénomène, valider ou invalider une hypothèse, appréhender tous les sens d'un mot, le linguiste ne peut plus se passer de telles ressources. Il doit disposer de corpus, lemmatisés et/ou annotés morphosyntaxiquement, et d'outils informatiques conviviaux lui permettant par exemple d'accéder aux entrées d'un dictionnaire à partir de la forme d'un mot, ou de rechercher des cooccurrences des mots d'un texte à partir des lemmes...

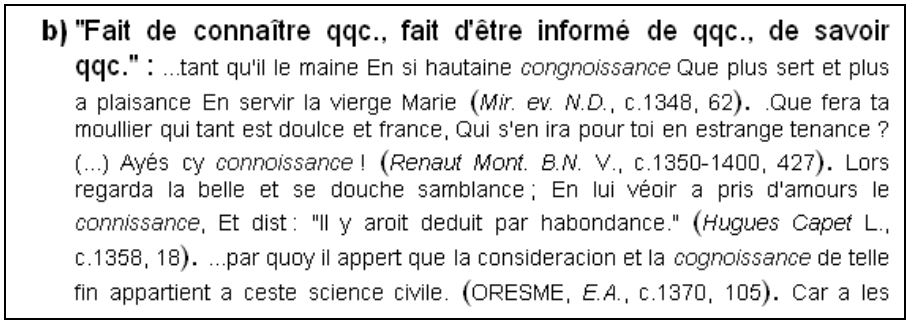
La lemmatisation consiste à fournir, pour une forme donnée, la représentation standardisée du mot correspondant, utilisée le plus souvent en entrée dans un dictionnaire de référence. On peut distinguer deux types de processus de lemmatisation : la lemmatisation hors contexte, lorsqu'on ne dispose que de la forme d'un mot sans contexte, c'est celle la plus utilisée pour consulter un dictionnaire, et la lemmatisation en contexte, lorsqu'on exploite le cotexte d'apparition d'une forme. Dans ce deuxième cas on pourra obtenir plus d'informations morphosyntaxiques, par exemple la catégorie grammaticale, et lever ainsi l'ambigüité sur les homographes. Ces aspects ont été relativement bien étudiés pour la langue moderne et de nombreux outils et ressources informatisés sont aujourd'hui disponibles pour faire de telles lemmatisations.

En ce qui concerne les états anciens de la langue, l'accès à de telles ressources est moins évident. Nous nous intéresserons plus particulièrement ici à un lemmatiseur hors contexte développé dans le cadre du projet Dictionnaire du Moyen Français. Le DMF est un dictionnaire électronique en ligne (www.atilf.fr/dmf). Il est constitué d'articles saisis directement en XML. Le dictionnaire est rédigé par étapes successives, à partir d'un corpus de 214 textes représentatifs de la langue (genre, auteur, œuvres…). L'étape actuelle, dont la fin est prévue en 2010, consiste à faire la synthèse des articles rédigés au cours des étapes précédentes (Martin, 1980, 2007). LGeRM (Lemmes, Graphies et Règles Morphologiques) a été conçu pour rendre plus conviviale la consultation du DMF. L'idée de base est que l'utilisateur fournit une forme quelconque rencontrée dans un texte, et que le dictionnaire lui propose la ou les entrée(s) correspondant à cette forme. Nous verrons que si cette idée est relativement simple à implémenter pour la langue moderne, il n'en est pas de même pour le français médiéval, compte tenu de ses spécificités.

1.2 *Spécificité du moyen français*

Le moyen français est la langue qui couvre la période allant du début du 14^{e} siècle au début du 15^{e}. Les dates du début et de la fin de cette période de la langue sont encore sujettes à discussion, mais la période retenue pour le DMF est 1330-1500 (Smith, 2002).

Il s'agit d'une langue en pleine évolution. Le moyen français présente déjà des aspects modernes mais conserve souvent des archaïsmes (*amy* ou *ami* ; *congnaistre ou conaitre*), l'orthographe des mots n'est pas encore stabilisée (*conaistre, conaitre* ou *connaitre*) et le système flexionnel est en pleine évolution (pluriel en s, x ou z : *ciels, cielx, cielz*). C'est aussi une période où de nouveaux mots apparaissent, par exemple dans les traductions en langue française des classiques grecs et latins. De plus il n'y a pas encore de français standard, et les textes sont parfois très marqués dialectalement par l'atelier de saisie du manuscrit : textes picards dans lesquels on trouve *chiel* pour *ciel*, ou textes anglo-normands où l'on rencontre *bastoun* pour *baston*). Un autre phénomène à prendre en compte relève des pratiques de transcription des manuscrits qui ont évolué au cours du temps. Aujourd'hui contrairement au siècle dernier, on modernise moins les formes, on segmente plus facilement les mots accolés, et les règles d'usage de la majuscule diffèrent.Regardons par exemple le mot *Connaissance* dans le DMF. En français moderne il possède les deux formes *connaissance* pour le singulier et *connaissances* pour le pluriel. Dans les exemples du DMF, on rencontre vingt formes différentes : *cognescence*, *cognissance*, *cognoiscence*, *cognoiscences*, *congneissance*, *connissance*… (voir la figure1). On voit l'extrême variation rencontrée dans le dictionnaire qui n'offre lui-même qu'un échantillon des graphies possibles du mot, on peut rencontrer d'autres formes dans le corpus de textes ayant permis de rédiger le dictionnaire (*congnoissanche*, *conaisanche*…). Cette liste de formes attestées dans le DMF ne peut être en aucun cas exhaustive. L'exploitation de nouveaux documents de cette époque non encore édités, nous amèneraient surement à rencontrer de nouvelles formes. Plus le mot comporte de syllabes, plus la variation est importante. Multiplié par le nombre d'entrées du dictionnaire, plus de 60 000 dans sa dernière version, on se rend bien compte qu'il est difficile voire impossible d'établir une liste exhaustive des formes possibles des mots.



b) "Fait de connaître qqc., fait d'être informé de qqc., de savoir qqc." : ...tant qu'il le maine En si hautaine *congnoissance* Que plus sert et plus a plaisance En servir la vierge Marie (*Mir. ev. N.D.*, c.1348, 62). .Que fera ta moullier qui tant est doulce et france, Qui s'en ira pour toi en estrange tenance ? (...) Ayés cy *connoissance* ! (*Renaut Mont. B.N.* V., c.1350-1400, 427). Lors regarda la belle et se douche samblance ; En lui véoir a pris d'amours le *connissance*, Et dist : "Il y aroit deduit par habondance." (*Hugues Capet* L., c.1358, 18). ...par quoy il appert que la consideracion et la *cognoissance* de telle fin appartient a ceste science civile. (ORESME, *E.A.*, c.1370, 105). Car a les

Figure 1 : *Un extrait de l'article* Connaissance *du DMF*

Le traitement informatique du moyen français devra tenir compte de ces deux faits : un grand nombre de possibilités pour une même forme morphosyntaxique et une liste complète des formes impossible à établir.

En ce qui concerne la consultation d'un dictionnaire électronique, l'utilisateur a en général la possibilité de parcourir la liste des entrées. *A priori* il ne devrait pas y avoir de différence entre la langue ancienne et la langue moderne, sauf qu'il n'y a pas de règle standard adoptée par l'ensemble de la communauté. Le *Godefroy* (Godefroy, 1881) et le *Tobler-Lommatzsch* (Tobler, 1925), les deux dictionnaires de référence pour l'époque médiévale, utilisent chacun leurs propres règles de lemmatisation. Le choix du DMF a été de moderniser les entrées, mais seulement pour les mots existant encore dans la langue moderne. Ainsi le mot *Physicien* (médecin) est traité sous *Fisiciien* dans le *Tobler-Lommatzsch* et sous *Fisicien* dans le *Godefroy*. Pour consulter les versions électroniques de ces dictionnaires il faut jongler entre plusieurs systèmes normatifs. Un dictionnaire électronique de base peut imposer à l'utilisateur de fournir l'entrée correspondant au mot qu'il cherche (cette entrée est en général le lemme et il convient alors d'enlever la marque du pluriel, du genre, la flexion verbale…) mais un dictionnaire plus convivial se doit de permettre un accès à partir de toute forme graphique du mot, il lui faut alors analyser cette forme et la rattacher à une ou éventuellement à plusieurs entrées. La technique utilisée en général consiste à s'appuyer sur l'ensemble des formes fléchies des mots : Morphalou (Morphalou, 2004) par exemple pour le TLFi (Dendien, 2002 ; ATILF, 2004). Mais cette technique ne peut pas être mise en œuvre pour un dictionnaire électronique convivial de la langue ancienne puisqu'il n'existe pas de liste exhaustive des formes fléchies. Ainsi, les versions électroniques du *Godefroy* (Godefroy, 2005) et du *Tobler-Lommatzsch* (Blumenthal, 2002) imposent de connaître l'entrée pour consulter un article.

En ce qui concerne la lemmatisation complète d'un texte, des études ont été entreprises en utilisant des outils d'analyse morphosyntaxique de type probabiliste tels que le TreeTagger (Schmid, 1994) ou le catégoriseur de Brill (Brill, 1992). Au départ il est toujours possible de construire un corpus annoté pour l'apprentissage. Néanmoins face à un nouveau texte comportant de nouvelles graphies, les outils ont des difficultés à attribuer un lemme aux graphies peu fréquentes (Prévost, 2000) (Kunstmann, 2006).

Conçu au départ pour faciliter la consultation du DMF, LGeRM (Lemmes, Graphies et Règles Morphologiques) tente de répondre à cette difficile question de la lemmatisation hors contexte en moyen français. Si l'utilisateur connait *a priori* le mot, il ne sait pas forcément quelle entrée a été retenue par le DMF, même si d'une manière générale, il est dit que les entrées ont été modernisées. S'il se trouve face à une forme qu'il n'arrive pas à associer à une entrée du dictionnaire, l'interface peut lui proposer une ou plusieurs entrées directement à partir de la forme. Ce cas de figure est d'autant plus fréquent que la connaissance de la langue médiévale par l'utilisateur est étendue ou non. Le DMF veut s'adresser non seulement aux spécialistes de la langue médiévale mais aussi à des utilisateurs moins savants. Face à la graphie *fusicien*, il n'est pas évident de penser immédiatement à l'entrée *Physicien* ; face à *polra* d'aller consulter le verbe *Pouvoir*.

Après cette introduction situant le contexte de nos travaux et la spécificité du moyen français, nous présenterons tout d'abord notre lemmatiseur, les matériaux à notre disposition pour le construire, l'architecture du système, les connaissances mises en œuvre et leur structuration, l'algorithme et enfin les résultats et les limites. Nous présenterons ensuite comment l'outil a été utilisé dans le DMF ou dans d'autres applications telles que la lemmatisation ou l'interrogation de textes, la construction automatique de glossaire.

2. LGeRM

2.1 *Les matériaux à notre disposition*

Les matériaux initiaux qui ont permis de réaliser la première version du lemmatiseur proviennent du Dictionnaire du Moyen Français. Le DMF est entièrement codé et exploité en XML. Les rédacteurs construisent leurs articles avec un éditeur XML. Les premiers articles avaient été saisis en 2001 au format SGML avec l'éditeur SoftQuad. L'éditeur utilisé actuellement est Corel XMetal. Le rédacteur est guidé dans la construction de son article et l'éditeur lui impose de baliser l'entrée, le code grammatical, des définitions, des exemples… et en particulier l'occurrence de l'entrée dans l'exemple (Souvay, 2004). Ce matériau souple a permis d'extraire deux listes sur lesquelles va s'appuyer le lemmatiseur : la liste des entrées du dictionnaire (lemmes) et la liste des occurrences de l'entrée (graphies) balisée dans l'exemple qui suit avec la balise <OCC>.

<P> <DISC> <NUM> **b)** </NUM> <DEF> "Fait de connaître qqc., fait d'être informé de qqc., de savoir qqc." </DEF> </DISC> <EXE> : <TEXTE> ...tant qu'il le maine En si hautaine <OCC> ***congnoissance*** </OCC> Que plus sert et plus a plaisance En servir la vierge Marie </TEXTE> <REF> (Mir. ev. N.D., c.1348, 62) </REF>. </EXE> <EXE> <TEXTE> .Que fera ta moullier qui tant est doulce et france, Qui s'en ira pour toi en estrange tenance ? (...) Ayés cy <OCC> ***connoissance*** </OCC> ! </TEXTE> <REF> (Renaut Mont. B.N. V., c.1350-1400, 427) </REF>. </EXE> <EXE> <TEXTE> Lors regarda la belle et se douche samblance ; En lui véoir a pris d'amours le <OCC> ***connissance*** </OCC>, Et dist : "Il y aroit deduit par habondance." </TEXTE> <REF> (Hugues Capet L., c.1358, 18) </REF>. </EXE> <EXE> <TEXTE> ...par quoy il appert que la consideracion et la <OCC> ***cognoissance*** </OCC> de telle fin appartient a ceste science civile. </TEXTE> <REF> (ORESME, E.A., c.1370, 105) </REF>. </EXE> </P>

Figure 2 : *Un extrait de l'article* Connaissance *du DMF au format XML*

Comme nous l'avons dit en introduction, la liste des graphies ne peut pas être exhaustive. Il faut donc trouver un moyen d'analyser les formes inconnues. Pour cela il va falloir mettre en œuvre des connaissances sur la morphologie de la langue médiévale. Il se trouve que dans le cadre d'un travail précédent, au cours d'un DEA (Souvay, 1986) nous avions mené une étude commune avec l'Équipe de l'Unité de Recherche sur le Français Ancien, Unité de Recherche Linguistique N 10. CNRS & Université Nancy 2. Les linguistes de cette équipe avaient acquis une expérience en

matière d'analyse de textes et ils avaient élaboré un modèle d'analyseur et fait l'inventaire des connaissances à utiliser (Monsonego, 1989). Ce travail de DEA a consisté à réaliser un premier prototype pour les mots invariables et le système nominal. Nous avions donc à notre disposition une centaine de règles morphologiques. L'algorithme actuel de LGeRM a été fortement inspiré par ce premier essai, mais ses principes sont plutôt opposés. Nous verrons plus loin, dans la partie algorithme, que LGeRM produit de nombreuses formes, alors que dans le travail précédent l'idée était de réduire les diverses formes d'un mot en une forme unique correspondant au lemme.

2.2 *Principe général*

Le système repose sur un algorithme qui va fournir une liste d'hypothèses de lemmes à partir d'une forme donnée en entrée. Son travail va consister à produire de nouvelles formes en faisant attention à ne pas boucler et à ne pas produire trop de résultats. Il s'appuie pour ses raisonnements sur une base de connaissances composée de trois éléments : la liste des lemmes, la liste des graphies et les règles de morphologiques.

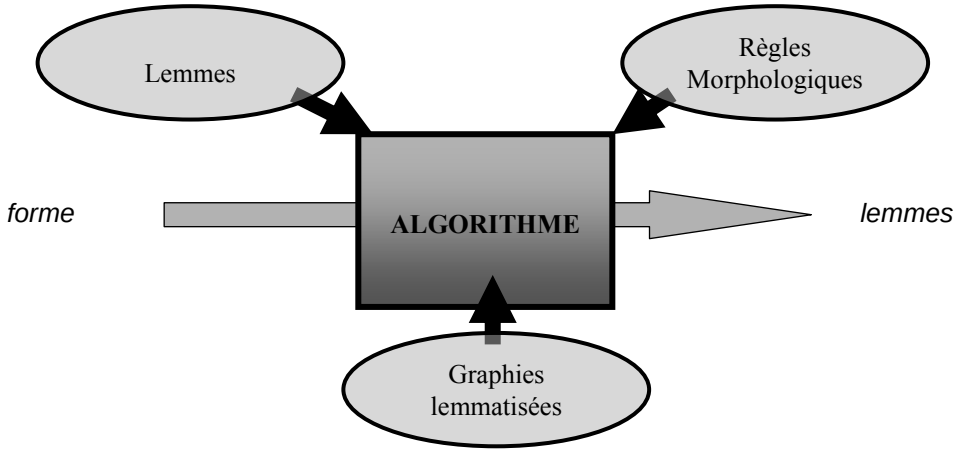


Figure 3 : *Architecture du système*

2.3 *Les ressources utilisées*

2.3.1. *Les lemmes*

La première version du lemmatiseur s'appuyait sur une version initiale du DMF mise en ligne en janvier 2003. Elle comportait environ 26 400 entrées. La dernière version du DMF en ligne depuis janvier 2009 a désormais une nomenclature d'un peu plus de 60 000 entrées, ce nombre ne devrait plus augmenter de manière

significative. La liste des lemmes a été mise à jour au fur et à mesure de l'évolution du DMF.

Les mots outils ne sont pas traités dans la version actuelle du DMF. Ils devraient faire l'objet d'une étude plus spécifique en 2010. Il manquait donc au lemmatiseur tous les lemmes correspondant à ces mots outils. Ils ont été rajoutés manuellement bien que ne correspondant à aucune entrée du DMF. Exemples de lemmes ajoutés : *à*, préposition ; *ains*, adverbe ; *le*, article ; *le*, pronom, etc.

2.3.2. *Les graphies exploitées*

Les matériaux lexicographiques du DMF ont permis d'extraire environ 80 000 graphies. Les premiers essais ont montré que la taille de la base de graphies semblait avoir déjà atteint un seuil critique suffisant pour fournir des résultats pertinents.

L'enrichissement s'est fait de plusieurs manières. Tout d'abord en générant automatiquement des formes au singulier à partir des pluriels ou inversement des formes au pluriel à partir des singuliers des substantifs et des adjectifs, en générant des formes féminines sur les adjectifs, etc.. Une seconde voie d'enrichissement a été d'extraire du DMF les références aux entrées des dictionnaires *Tobler-Lommatzsch* et *Godefroy* cités dans les articles. Cela a permis d'enrichir la base en graphies plus anciennes que celles traitées par le DMF et ainsi d'assurer, d'une part, une certaine continuité diachronique des mots, et, d'autre part, de satisfaire les utilisateurs plus habitués à la forme normative différente de ces dictionnaires. La taille de la base est ainsi passée à 150 000 en avril 2004.

Le nombre de formes dans la base de graphies n'a pas de conséquence réelle sur les pertinences des réponses. Il apparaît que l'interaction entre règles morphologiques et graphies lemmatisées est plus importante. Quand il n'y avait que 80 000 graphies le système fonctionnait déjà correctement. L'ajout de règles morphologiques a permis de résoudre les cas qui n'aboutissaient pas ou les cas avec résultat erroné. Néanmoins l'ajout de graphies permet au système de répondre plus rapidement (il y a moins d'hypothèses à générer et explorer) et plus spécifiquement sur les mots courts qui peuvent produire des lemmes aberrants. Par exemple la graphie *peu* serait reconnue non seulement comme une forme de *Pouvoir*, verbe et *Peu*, adverbe mais aussi, par application des règles, à *Pieu*, *Peau*, *Pal*, *Pouls*, *Pouce*, *Poil* et *Pou*.

Pour poursuivre l'enrichissement de la base, notamment en ce qui concerne les formes flexionnelles verbales mal représentées dans les exemples du DMF, la nomenclature du Trésor de la langue Française (TLF) et sa liste de 400 000 formes fléchies a été, elle aussi, exploitée. En appliquant des règles graphiques ou morphologiques sur les formes modernes, pour les lemmes déjà présents dans le DMF, un programme a généré des formes anciennes. Ainsi pour le substantif *Abattement* par exemple, on a produit les formes *abattemens*, *abattements*, *abbattementz* et *abbattemenz*. Pour les verbes du premier groupe des formes en finale *-eraye*, *-erayes*, *-eroie*, *-eroies*, etc. 250 000 formes nouvelles ont ainsi été

ajoutées. Parmi ces 250 000 nouvelles formes environ 115 000 étaient attestées dans les bases textuelles médiévales de l'ATILF. Au final en décembre 2004, il y avait 400 000 éléments dans la liste des formes.

Une autre source d'enrichissement pour la flexion verbale a été le fonds de formes flexionnelles établi, dans les années soixante, par Robert Martin. Il s'agit de formes verbales analysées, pour la période allant de l'ancien français au français de la Renaissance présentes dans 41 dictionnaires. Ce fonds de 116 000 formes est valorisé sur le site de l'ATILF à l'adresse www.atilf.fr/bgv. Ce fonds est intéressant et très complémentaire car il comporte de nombreuses formes irrégulières de verbes courants qui ne pouvaient pas être générées automatiquement, par exemple *unt*, *oi*, *oc*, *orent*, *averai* pour le verbe *Avoir*. Avec l'ajout de nouveaux lemmes traités dans les versions successives du DMF, l'exploitation plus fine des textes existants et l'introduction de nouveaux textes, l'intégration partielle de la base de graphies verbales, la base contient aujourd'hui près de 650 000 formes.

2.3.3. *Les règles graphiques et morphologiques*

2.3.3.1 Définition

La définition des règles graphiques et morphologiques s'appuie au départ sur des travaux réalisés en 1986 par l'Équipe de l'Unité de Recherche sur le Français Ancien, (Souvay, 1986). Les règles initiales ont été adaptées, il ne s'agit plus de ramener au lemme mais de produire des formes alternatives. Il a fallu ajouter les règles pour la flexion verbale qui ne faisait pas l'objet de l'étude. Des ouvrages sur le système flexionnel médiéval (Buridant, 2000) et la base de graphies verbales ont permis d'établir les règles, qui ont été ensuite validées en s'appuyant sur les ressources informatisées du DMF (dictionnaire et textes).

Ces règles sont des règles de réécriture sur les caractères de la forme en cours de traitement. Chaque règle possède une précondition facultative et une postcondition elle aussi facultative. La précondition porte sur l'initiale du mot, sa finale, ou le contexte entourant un groupe de lettres en exploitant des fonctions de type `entre`, `suivi de`, `non suivi de`, `précédé de`, `précédé de sauf`. On peut y faire référence à un caractère ou à une séquence de caractères. Ces caractères peuvent être eux-mêmes définis par une liste extensive ou par des listes prédéfinies telles voyelles ou consonnes. La postcondition quant à elle porte sur le résultat : on peut indiquer si la règle porte ou non sur le système verbal, si la règle porte sur les noms propres, ou encore si la règle a conduit à un succès (la forme produite est dans la base de graphies).

Quatre grandes catégories de règles sont présentes dans le système : des règles morphologiques sur la flexion verbale, des règles morphologiques sur la flexion nominale ou adjectivale, des règles d'agglutination et des règles générales purement orthographiques.

2.3.3.2 La flexion verbale

Les règles sur la flexion verbale portent systématiquement sur la finale des mots. Elles sont regroupées en fonction d'un infinitif générique pour les dérivés de verbes (verbes du premier groupe, verbes dérivés de *Mettre*, verbes en finale *-uire*…) ou rassemblées en fonction de règles plus générales communes à tous les verbes (participe présent, participe passé, altération de la flexion…).

Les règles sur le système verbal sont les plus nombreuses, il en existe environ quatre mille. Parmi ces règles on peut, entre autre, distinguer celles qui :

(a) tentent de ramener la graphie à l'infinitif :

Précondition :	en finale
Règle :	MECT → METTRE
Postcondition :	est verbal
Exemple :	*admect* → *admettre*

(b) tentent de moderniser une forme ancienne pour la ramener à une forme moderne connue :

Précondition :	en finale
Règle :	TERAIT → TRAIT
Postcondition :	est verbal
Exemple :	*transmetterait* → *transmettrait*

(c) tentent de passer à une autre personne de la conjugaison :

Précondition :	en finale
Règle :	OIENT → OIT
Postcondition :	est verbal
Exemple :	*escorceroient* → *escorceroit*

2.3.3.3 La flexion nominale et adjectivale

Les règles sur la flexion nominale et adjectivale portent sur la finale des mots comme pour le système verbal. Les règles sont regroupées en fonction de la finale du lemme (finale en -AL, finale en -EUX…). Il existe environ deux cents règles sur le système flexionnel nominal et adjectival. Parmi ces règles, il y a systématiquement celle qui permet d'obtenir la forme masculin singulier :

Précondition :	en finale
Règle :	ALES → AL
Postcondition :	est nominal ou adjectival
Exemple :	*abbaciales* → *abbacial*

Il existe des règles qui permettent de passer sur une variante graphique sans pour autant essayer de trouver la forme correspondant au lemme :

Précondition :	en finale
Règle :	ÉZ→ ETZ
Postcondition :	est nominal ou adjectival
Exemple :	*coqueléz* → *coqueletz*

Il existe aussi des règles pour moderniser les formes anciennes :

Précondition :	en finale
Règle :	ALZ → AUX
Postcondition :	est nominal ou adjectival
Exemple :	*deloyalz* → *deloyaux*

2.3.3.4 Les règles d'agglutination

Dans les textes médiévaux, beaucoup de mots sont agglutinés. Il peut s'agir de l'article, d'un adverbe, d'un pronom, d'un élément formant : de telles agglutinations sont assez nombreuses avec les mots *arrière, avant, beau, bien, contre, par, se, sur, très, vice*… Les règles vont faire en sorte que ne soit conservé que le mot plein. Voici à titre d'exemple un court extrait faisant apparaître une agglutination, tiré de Jean Regnier, *Les Fortunes et adversitez*, 1432-c.1465 :

> *Je l'endure* tresdebonnairement *En gré je prens tout son commandement, Ne plorez plus, prenez ce qu'il advient, En attendant de Dieu le jugement.*

Une centaine de règles sont nécessaires pour gérer la plupart de ces cas de figure.

2.3.3.5 Les règles générales

Les règles générales se répartissent en sous-catégories très diverses. Elles recoupent parfois certaines règles sur la flexion. Les recoupements seront gérés par l'algorithme au moment de la vérification de la postcondition.

Il existe des règles de modernisation ou de vieillissement des graphies :

Précondition :	
Règle :	Y → I
Postcondition :	
Exemple :	*gay* → *gai*
Précondition :	en initiale et suivi de voyelle
Règle :	SÇ→ S
Postcondition :	
Exemple	*presçavoir* → *presavoir*

Il existe des règles spécifiques à certains régionalismes, par exemple pour l'anglo-normand, le français parlé/écrit en Angleterre à l'époque médiévale :

Précondition :
Règle : OUN→ ON
Postcondition :
Exemple : *bastoun* → *baston*

Il existe des règles inspirées de la phonétique ayant de nombreuses attestations dans le corpus (les manuscrits à transcrire peuvent être dictés dans les ateliers) :

Précondition : suivi de [E,I]
Règle : C→SS
Postcondition
Exemple : *abaicement* → *abaissement*

Précondition : En initiale
Règle : H→
Postcondition :
Exemple : *habandon* → *abandon*

Il existe environ quatre cents règles d'ordre général.

2.3.3.6 La représentation des règles

Afin de faciliter leur saisie, leur mise au point et leur relecture, et aussi afin d'assurer leur portabilité future pour d'autres utilisations, les règles ne sont pas codées directement dans le programme informatique. Elles sont représentées en utilisant le format XML. Cela garantit une indépendance par rapport à l'algorithme et permet de modifier le système à l'aide d'un simple éditeur XML. Elles sont ainsi plus abordables par un non informaticien. Elles sont affichées en utilisant une feuille de style qui reproduit le formalisme initial des règles tel qu'il avait été défini par l'équipe du CRAL et qui semble mieux parler au linguiste. Une règle est de la forme si une condition est remplie alors on effectue une action. La condition regroupe la précondition et la postcondition, et l'action contient la règle de réécriture. Ainsi une des règles de décodage du participe présent pour un verbe du premier groupe s'écrit dans ce formalisme :

si (en finale) et (verbal) alors ANS → ER finsi

Elle correspond à la règle suivante :

Précondition : en finale
Règle : ANS → ER
Postcondition : est verbal

2.4. *Algorithme*

Le principe général de l'algorithme n'est pas, comme dans une lemmatisation classique, d'essayer de trouver la forme normalisée du lemme, mais de trouver une forme connue dans la base de graphies la plus proche possible de la forme à lemmatiser.

Au départ, si la forme à lemmatiser est dans la base de graphies, le lemmatiseur propose les lemmes attachés à cette forme, la lemmatisation est alors terminée. Si au contraire, elle n'est pas dans la base de graphie, le lemmatiseur applique toutes les règles morphologiques sur cette forme inconnue. Si parmi les formes générées dans l'itération précédente il existe une forme connue, l'algorithme s'arrête. S'il n'existe aucune forme connue, le processus est réappliqué sur chacune des formes générées. L'algorithme s'arrête si aucune forme nouvelle n'est produite ou si le nombre de formes générées dépasse un seuil maximum de formes autorisées.

Il existe un mode débrayé dans lequel au départ le lemmatiseur ne vérifie pas que la forme initiale est connue, cela permet de proposer des lemmes supplémentaires sur des formes présentes dans la base de graphies.

La figure 5 présente le cœur de l'algorithme. Les variables sont mises en italique, elles portent des noms explicites. Les variables principales sont les listes *BaseDeGraphies*, *BaseDeRègles*, *FormesProduites*, *Hypothèses*, la chaîne *FormeÀLemmatiser*, l'entier *NiveauDeProfondeur*, et le booléen *Continuer*. L'affectation d'une variable est représentée par la flèche ←.

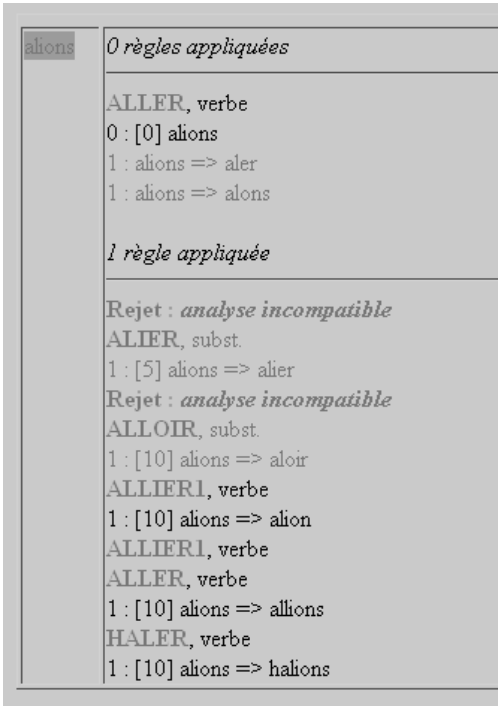
```
si (Appartient(FormeÀLemmatiser, BaseDeGraphies) et mode_débrayé_non_activé) alors
    Hypothèses ← ExtraireHypothèses(FormeÀLemmatiser, BaseDeGraphies) ;
sinon
    /*** initialisations ***/
    NiveauDeProfondeur← 0 ; ya_succes[NiveauDeProfondeur] ← FAUX ;
    FormesProduites ← Ajouter(FormeÀLemmatiser, NiveauDeProfondeur) ;
    Incrémente (NiveauDeProfondeur) ;
    ya_succes[NiveauDeProfondeur] ← FAUX ;
    forme ← Tête(FormesProduites, NiveauDeProfondeur -1) ;
    Continuer ← VRAI ;
    /*** itération ***/
    tantque (Continuer) faire
        /*** parcours des règles ***/
        regle ← Tête(BaseDeRègles) ;
        tantque (regle ≠ NULL) faire
            nouvelle_forme ← AppliquerRegle(forme, regle) ;
            si nouvelle_forme ≠ NULL alors
                si Appartient(nouvelle_forme, BaseDeGraphies) alors
                    ya_succes[NiveauDeProfondeur] ← VRAI ;
                    succes ← ExtraireHypothèses(nouvelle_forme, BaseDeGraphies) ;
                    Hypothèses ← Ajouter(succes) ;
                finsi
                FormesProduites ← Ajouter(nouvelle_forme, NiveauDeProfondeur) ;
            finsi
            regle ← RegleSuivante(BaseDeRègles) ;
        fin tantque
        forme ← FormeSuivante(FormesProduites, NiveauDeProfondeur-1) ;
        /*** calcul du test d'arrêt ***/
        si ((forme = NULL) ou ya_succes[NiveauDeProfondeur-1]) alors
            Continuer ← FAUX ;
        sinon
            si (Cardinal(FormesProduites) < seuil_arret_succes) alors
                Incrémente (NiveauDeProfondeur) ;
                forme ← FormeSuivante(FormesProduites, NiveauDeProfondeur-1) ;
                encore ← (forme ≠ NULL) ;
            sinon
                Continuer ← FAUX ;
            finsi
        finsi
        si (Cardinal(FormesProduites) > seuil_arret_urgence) alors
            Continuer ← FAUX ;
        finsi
    fin tantque
finsi
```

Figure 4 : *Cœur de l'algorithme mis en œuvre pour lemmatiser une forme*

Voyons comment fonctionne LGeRM sur un exemple. Considérons la forme *alions*. C'est une forme connue de la base de graphies, il s'agit d'une forme du verbe *Aller* :

Figure 5 : *Lemmatisation par défaut de « alions »*

Passons en mode débrayé et regardons le résultat de LGeRM pour la même forme :



Détail des règles appliquées

R1_11-02(alions, I)=>alyons
R1_21-03(alions, S)=>alionz
Chanter[IPr1P](alions, ONS$)=>**alier**
Chanter[IIm1P](alions, IONS$)=>**aler**
Chanter[SPr1P](alions, IONS$)=>**aler**
Valoir[IIm1P](alions, ALIONS$)=>**aloir**
Rendre[IPr1S](alions, ONS$)=>aliondre
TransformerFlexion(alions, IONS$)=>alir
TransformerFlexion(alions, IONS$)=>**alons**
AuSingulier(alions, S$)=>**alion**
R2_15-05(alions, L)=>**allions**
R2_27-02(alions, A)=>adlions
R2_33-02(alions, A)=>**halions**

Les formes connues sont affichées en gras.

Figure 6 : *Lemmatisation de « alions » en mode débrayé*

Dans la première itération LGeRM propose *Aller* à partir de la forme *alions* sans appliquer de règle. On remarque en grisé qu'il a obtenu la même solution pour *aler* en appliquant la règle IONS → ER et *alons* en appliquant IONS → ONS.
Dans la seconde itération LGeRM propose quatre nouveaux lemmes. *Alier* et *Alloir* ont été rejetés car la postcondition 'est verbal' n'est pas respectée.

Allier a été obtenu à partir de la graphie *alion* (suppressions du S final, variante graphique pour les verbes du premier groupe, indicatif présent première personne du pluriel), et *Haler* à partir de la forme *halions* (ajout d'un H à l'initiale).

2.5. *Résultats et limites*

En mode consultation de dictionnaire, les résultats sont satisfaisants. L'utilisateur est, dans la très grande majorité des cas, guidé vers la bonne entrée. En cas d'homographie, on lui donne la possibilité d'afficher les différents articles pour se faire une idée du sens du mot. Les cas d'erreur proviennent des homographes, pour lesquels seule la graphie d'un des homographes est connue : dans ce cas le système ne propose pas l'autre l'entrée. Par exemple, dans le cas de la forme *alions*, le DMF ne propose pas le verbe *Allier* car cette attestation n'est pas dans la base de graphies. Pour l'obtenir il faut demander au dictionnaire d'émettre plus d'hypothèses (mode débrayé).

En mode lemmatisation de textes, les hypothèses multiples sont une gêne et produisent beaucoup de bruit. Un homographe très rare d'un mot est systématiquement proposé. Pour la forme *a* par exemple, le système propose quatre hypothèses : *A*, substantif *Ah*, interjection *Avoir1*, verbe et *À*, préposition alors que *A*, substantif et *Ah* interjection sont très rares. Le système ne lève pas non plus l'ambigüité entre le verbe et la préposition. Mais il faut relativiser, ce phénomène se produit pour un petit nombre de formes relativement fréquentes. Pour la plupart des formes, plus particulièrement les graphies rares, le système est très pertinent dans ses analyses. Ce phénomène n'est cependant pas spécifique au moyen français.

Une première évaluation du lemmatiseur a été réalisée en 2006 sur un texte du corpus DMF : Jacques LEGRAND, *Livre de bonnes meurs*, 1410 (Souvay, 2007). Le texte comporte 46 153 mots :

- dans 60% des cas le lemmatiseur fournit un lemme unique et ce lemme est correct,
- dans 39% des cas il donne plusieurs lemmes et le lemme correct est dans les hypothèses,
- reste 1% d'erreurs, qui se produisent majoritairement sur des noms propres homographes d'un nom commun.

En ce qui concerne les erreurs, dans seize cas le lemmatiseur n'a pas été capable de formuler d'hypothèse. Six cas portaient sur des adjectifs numériques non prévus *.XXX Ie. .C VIIIe* (en fait l'espace à l'intérieur du nombre n'est pas autorisé). Des graphies étonnantes ont été rencontrées. S'agit-il d'une erreur de transcription pour *vraincre*, d'une abréviation ou d'une erreur de transcription pour *entendment*, d'une erreur de numérisation pour *finablemlent* ? Nous n'avons pas eu la possibilité de retourner au manuscrit d'origine pour le vérifier. La forme *entendment* nous a conduit à regarder de plus près des cas semblables dans le corpus et finalement d'ajouter de nouvelles règles morphologiques. Le cas *olyfamble* correspond à une variante inconnue du mot *oriflamme* et difficilement atteignable en utilisant des règles de réécriture. Deux agglutinations de *vaine* et *gloire* pour la forme *vainegloire*

sont non prévisibles, une transcription moderne découperait sans doute les deux mots. Un mot nouveau *tripartite* attesté deux fois a été découvert et introduit depuis dans le DMF.

Il est à noter que le texte n'avait pas été spécialement préparé pour une lemmatisation, une transcription plus récente du manuscrit fournirait par exemple un balisage des noms propres, ce qui améliorerait notablement les résultats. Chaque texte est un cas particulier, une bonne préparation ne peut qu'améliorer les résultats. Un second passage du lemmatiseur après correction éventuelle du texte et enrichissement de la base de connaissances (ajout de graphies, nouvelles règles morphologiques) est aussi recommandé.

2.6. *Perspectives*

Les perspectives d'évolution du lemmatiseur sont plus ouvertes du coté de la lemmatisation de textes que de la consultation du dictionnaire. En effet dans ce dernier cas, les objectifs sont atteints. Il faudrait éventuellement réviser à moyen terme le contenu de la base de graphies : les enrichissements successifs du DMF, décalés dans le temps, ont pu faire apparaître certaines formes non encore intégrées dans la base de graphies, plus particulièrement au niveau des homographes.

La perspective la plus intéressante se situe au niveau de la levée de l'ambigüité des lemmes en cas de lemmatisation d'un texte. À la lecture, on distingue très bien la différence entre le substantif *Livre* et le verbe *Livrer* pour la graphie *livre*. Une étude contextuelle du voisinage pourrait être envisagée. Une autre direction serait de combiner LGeRM avec un analyseur morphosyntaxique. Cette approche n'a pu être mise en œuvre pour l'instant, faute de projet support et surtout de textes lemmatisés et validés pour l'apprentissage. Un moment envisagée pour le projet *Modéliser le changement, les voies du français*, elle n'a pu aboutir faute de temps et de disponibilité des chercheurs concernés (Martineau, 2007). Il s'agissait d'utiliser TreeTagger qui est capable de mettre une étiquette morphologique en utilisant des critères probabilistes. Mais il ne peut lemmatiser des formes inconnues. LGeRM fait des propositions pertinentes sur les formes inconnues, mais n'est pas capable de lever l'ambiguïté entre un substantif, un verbe ou une interjection. La complémentarité des approches pourrait être fructueuse.

Le lemmatiseur pourrait être utilisé pour une autre langue latine médiévale où les phénomènes de variabilités des graphies sont identiques. L'adaptation serait possible pour peu que l'on dispose d'une base de graphies suffisamment riche et au prix d'une adaptation des règles morphologiques plus particulièrement au niveau de la flexion.

3. Utilisations du lemmatiseur

3.1. *DMF*

Le lemmatiseur est mis en œuvre de manière transparente dans l'interface du DMF (www.atilf.fr/dmf) pour la recherche sur les entrées. L'utilisateur introduit une forme et le DMF indique les entrées possibles (lemmes) pour cette forme.

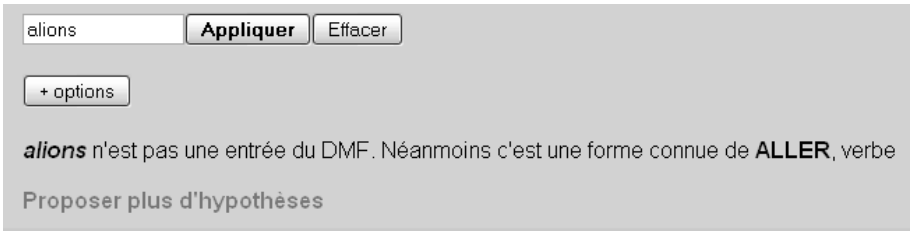


Figure 7 : *Recherche de l'entrée « Alions » dans le DMF*

Lors de l'affichage d'un article, il est possible d'enclencher une hyper navigation à partir des exemples par un clic sur un mot des exemples, qui provoque l'appel du lemmatiseur et propose alors les entrées possibles pour le mot.

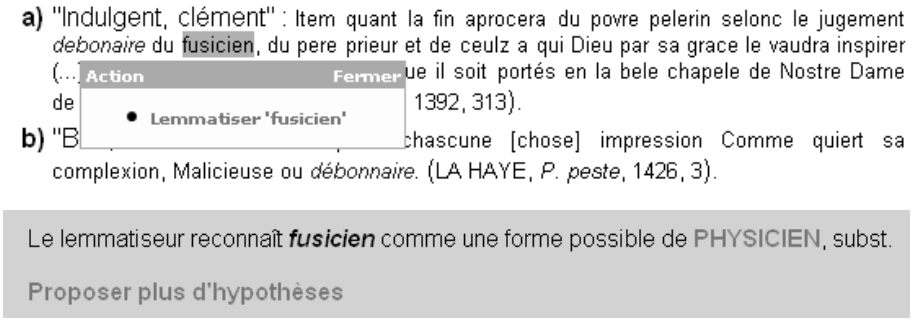


Figure 8 : *Lemmatiser la forme « fusisien »*

Le lemmatiseur est un outil accessible en ligne via le DMF (www.atilf.fr/dmf). Dans le menu lemmatisation, un formulaire permet d'étudier un mot, un extrait de texte ou même un texte balisé au format TEI (www.tei-c.org/). L'algorithme peut se paramétrer avec les options de traitement et les options avancées. L'affichage du résultat peut se paramétrer grâce aux options d'affichage et type de lemme.

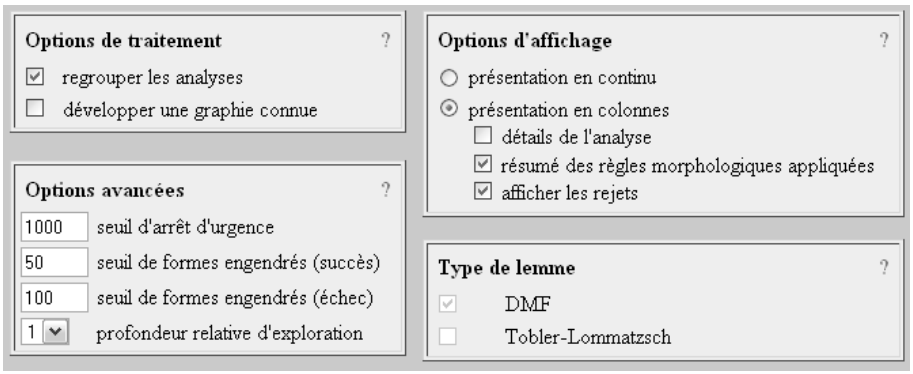


Figure 9 : *Paramétrage du lemmatiseur en ligne*

Pour chaque mot analysé on peut suivre le raisonnement du lemmatiseur, en particulier les règles appliquées et la liste des formes produites (cf. figure 7 ci-dessus).

Nous disposons aussi d'une version du lemmatiseur sous forme d'exécutable sous le système d'exploitation Microsoft Windows. Elle est accessible dans le cadre d'une collaboration avec le laboratoire ATILF.

3.2. *Autres dictionnaires*

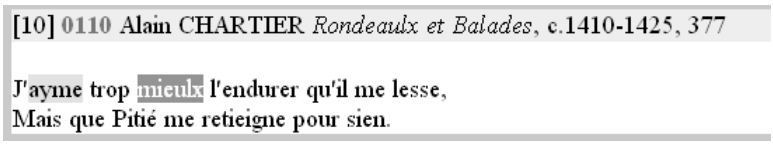
Le lemmatiseur est utilisé par le Dictionnaire Électronique de Chrétien de Troyes (Kunstmann, 2007) pour gérer de manière transparente, comme pour le DMF, la recherche sur les entrées. Le dictionnaire étant normé selon les entrées du *Tobler-Lommatzsch* (ancien français), le lemmatiseur a été adapté. Cela rend LGeRM compatible avec les outils et projets travaillant sur l'ancien français.

Figure 10 : *Recherche de l'entrée Aimer dans le DÉCT*

3.3. *Interrogation de textes non lemmatisés*

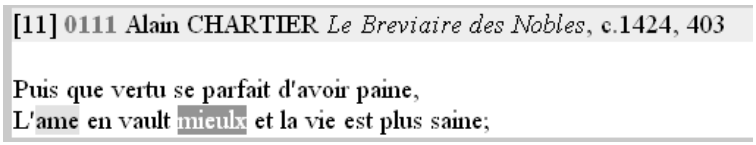
L'interrogation d'un corpus de textes médiévaux est une opération difficile si les textes ne sont pas lemmatisés. La variation graphique est telle qu'on ne peut pas imaginer toutes les formes que va prendre un mot. Le corpus DMF n'échappait pas à cette règle au départ. Géré via FRANTEXT (Bernard 2002), seule la flexion moderne était accessible en utilisant les expressions Stella (langage spécifique à FRANTEXT). Il restait à l'utilisateur à se créer des listes de mots, avec toute l'incertitude sur la complétude de sa liste. La lemmatisation du corpus de 218 textes n'étant pas envisageable, la solution pour l'interrogation par lemme a été d'utiliser la base de connaissances du lemmatiseur et plus particulièrement la base de graphies. Avec une interface adaptée, il est aisé de rechercher toutes les formes d'un lemme ou encore la cooccurrence de deux mots. Le corpus DMF est le premier corpus en langue médiévale pouvant être interrogé de la sorte. Toutefois le problème des homographies produit du bruit, par exemple le moteur de recherche est incapable de distinguer dans le texte si la forme *ame* est une forme du lemme *Aimer* ou du lemme *Âme*. Un autre cas de figure est lié au fait que la base de graphies présentant des lacunes, le moteur de recherche peut parfois ne pas détecter des attestations d'un lemme dans les textes dans la mesure où l'ensemble des formes du corpus des textes supports du DMF, de même que l'ensemble des homographes ne sont pas dans la base de graphies.

Voici un exemple de résultat correct de la recherche de *Aimer* et *Mieux* en cooccurrence :


[10] 0110 Alain CHARTIER *Rondeaulx et Balades*, c.1410-1425, 377

J'ayme trop mieulx l'endurer qu'il me lesse,
Mais que Pitié me retieigne pour sien.

Voici un exemple de résultat erroné, *ame* est ici une forme du substantif *Âme* et pas du verbe *Aimer* :


[11] 0111 Alain CHARTIER *Le Breviaire des Nobles*, c.1424, 403

Puis que vertu se parfait d'avoir paine,
L'ame en vault mieulx et la vie est plus saine;

3.4. *Édition électronique*

Le DMF offre la possibilité de faire un lien direct sur ses articles depuis n'importe quelle application développée sur la toile. La transcription d'un manuscrit est un travail lourd, même si la TEI a contribué à harmoniser les pratiques. Construire le glossaire correspondant constitue un travail supplémentaire qui n'est pas forcément la motivation première des équipes. L'exploitation du DMF, qui est la

référence pour cette période de la langue, permet de l'éviter en générant un simple lien de la forme www.atilf.fr/dmf/definition/<lemme> qui permet d'afficher l'article <*Lemme*> du DMF. Si le texte n'est pas lemmatisé, il convient alors d'exploiter LGeRM afin d'analyser la forme, en générant cette fois-ci un lien de la forme www.atilf.fr/dmf/morphologie/<forme>.

Ainsi www.atilf.fr/dmf/definition/pouvoir permet d'afficher l'article *Pouvoir* et www.atilf.fr/dmf/morphologie/polra d'étudier la forme *polra*.

3.5. *Glossaire en ligne*

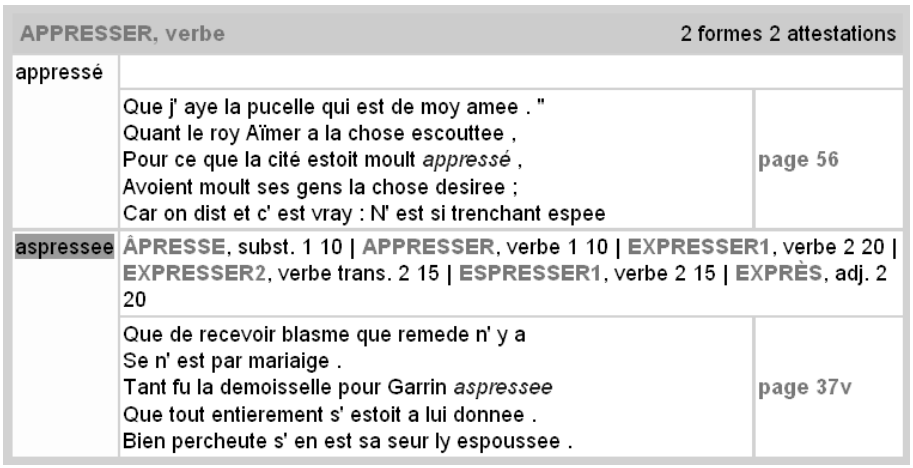
Une autre voie, qui a été explorée pour l'utilisation du lemmatiseur, est la construction automatique du glossaire d'un texte. Tous les mots d'un texte encodé en respectant les recommandations de la TEI sont passés au lemmatiseur et ensuite un traitement regroupe les formes par lemmes. On obtient alors une concordance au niveau des lemmes et non plus au niveau des formes comme les logiciels existant le pratiquent jusqu'à présent. Cet outil est en cours de développement dans le cadre d'un projet commun « *Le Dictionnaire de moyen français et autres dictionnaires de langues vernaculaires médiévales : principes, méthodologie, pratique, problèmes et solutions* » financé par le CNRS et la British Academy. Ce projet regroupe l'Université d'Edimbourg, l'Université de Sheffield, l'Université de Liverpool et l'ATILF pour un travail sur des transcriptions de Christine de Pizan, manuscrit de Harley et sur les *Chroniques* de Froissart. Une version de démonstration pour un texte libre de droits est en accès libre sur le site du DMF dans le menu Lemmatisation. Il convient pour cela de procéder en deux temps :

(1) Un premier passage pour nettoyer le texte. Le lemmatiseur met l'accent sur des erreurs d'encodage du texte XML ou des erreurs de transcriptions de mots. Ils montrent aussi l'importance de l'encodage des noms propres. Des fonctionnalités permettent d'accéder facilement aux mots non reconnus, aux mots où le lemmatiseur a des incertitudes sur l'analyse…

(2) Un second passage est alors possible, après enrichissement éventuel de la base de graphies. Les regroupements par lemmes sont pertinents, les formes inconnues sont bien regroupées. Il reste du bruit sur les homographes au niveau des lemmes et des noms propres si ces derniers ne sont pas encodés.

Les premiers résultats sont intéressants : l'outil permet en effet d'assurer automatiquement la lemmatisation de l'ensemble des formes du texte, lemmatisation attestée pour les lemmes connus du DMF, ou hypothèse de lemmatisation à valider linguistiquement pour les formes inconnues. Ainsi cet outil permet de détecter de nouveaux mots candidats à enrichir le DMF dans une version ultérieure.

Dans l'exemple qui suit le lemmatiseur propose le lemme *Apresser* pour deux formes du texte *appressé*, forme déjà connue, et *aspressee*, forme inconnue. Pour cette dernière, il envisage les deux lemmes *Apresse* et *Apresser* en appliquant une règle et les quatre lemmes *Expresser1*, *Expresser2*, *Espresser* et *Exprès* en appliquant deux règles.



APPRESSER, verbe	2 formes 2 attestations	
appressé		
	Que j' aye la pucelle qui est de moy amee . " Quant le roy Aïmer a la chose escouttee , Pour ce que la cité estoit moult *appressé* , Avoient moult ses gens la chose desiree ; Car on dist et c' est vray : N' est si trenchant espee	page 56
aspressee	ÂPRESSE, subst. 1 10 | APPRESSER, verbe 1 10 | EXPRESSER1, verbe 2 20 | EXPRESSER2, verbe trans. 2 15 | ESPRESSER1, verbe 2 15 | EXPRÈS, adj. 2 20	
	Que de recevoir blasme que remede n' y a Se n' est par mariaige . Tant fu la demoisselle pour Garrin *aspressee* Que tout entierement s' estoit a lui donnee . Bien percheute s' en est sa seur ly espoussee .	page 37v

Figure 11 : *Entrée* Appresser *du glossaire*

4. Conclusion

Dans cet article nous avons présenté un outil de lemmatisation des mots du français médiéval plus particulièrement focalisé sur la période du moyen français (1330-1500). Nous avons présenté les difficultés spécifiques au traitement automatique des mots médiévaux et montré en quoi les techniques de lemmatisation traditionnelles mises en œuvre pour la langue moderne ne fonctionnent pas pour le moyen français. Dans le contexte plus particulier du projet du DMF, nous avons proposé une solution adaptée à la consultation conviviale du dictionnaire fondée sur le lemmatiseur LGeRM. Ce lemmatiseur a suscité beaucoup d'intérêt dans la communauté des médiévistes, ce qui a conduit à l'intégrer dans différentes applications autres que l'accès à un dictionnaire telles : la lemmatisation complète et l'interrogation de textes, l'aide à l'édition électronique de textes en moyen français ou la construction automatique de glossaires. Défini au départ spécifiquement pour le moyen français, il a montré sa robustesse pour d'autres périodes de la langue, comme l'atteste son exploitation dans le cadre du projet de Dictionnaire Electronique de Chrétien de Troyes (ancien français), et du projet de l'ATILF d'étude du français préclassique.

L'outil est accessible sur le web à l'adresse www.atilf.fr/dmf, une interface permet de paramétrer le logiciel, de tester ses performances et de suivre son raisonnement.

Bibliographie

ATILF *Trésor de la Langue Française informatisé*, CNRS Editions, 591 p. et CD du texte intégral, Version PC, ISBN 2-271-06273-X, 2004, Version Mac OX X, ISBN 2-271-06365-5, 2005.

Bernard P., Dendien J., Lecomte J., Pierrel J.-M. (2002), « Un ensemble de ressources informatisées et intégrées pour l'étude du français : frantext, tlfi, dictionnaires de l'Académie et logiciel stella, présentation et apprentissage de leurs exploitations », *Actes de TALN 2002*, vol 2, p. 3-36, Nancy, 24-27 juin 2002.

Blumenthal P., Stein A. (2002), *Altfranzösiches Wörterbuch von Tobler, Lommatzsch*, Franz Steiner Verlag Stuttgart.

Brill. E. (1992), « A simple rule-based part of speech tagger ». *In Proceedings of the Third Conference on Applied Computational Language Processing*, p. 178–185, Trento.

Buridant C. (2000), *Grammaire nouvelle de l'ancien français*. Paris : Sedes.

Dendien J., Pierrel J.M. (2003), « Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », *TAL* Vol 44 – n° 2/2003, Hermes Sciences Edition, p. 11-37.

Godefroy F. (1881), *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle*, Paris F. Vieweg, Emile Bouillon, 10 tomes, 1881-1902.

Godefroy F. (2005), Version électronique du *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle*, Champions électronique, Paris.

Martin R. (1980), « Pour un dictionnaire du Moyen Français ». *In : Du Mot au Texte. Actes du IIIe Colloque international sur le Moyen Français*, Düsseldorf 1980. P. Wunderli, Tübingen, G. Narr, 1982.

Martin R., Gerner H., Souvay G. (2007), « Présentation de la seconde version du DMF ». *Actes de CILPR 2007 Congrès International de Linguistique et de Philologie Romane*, Innsbruck, Autriche, 3-8 septembre 2007. A paraître.

Martineau F. (2007), « Le pouls du changement : le projet Modéliser le changement : les voies du français », *International Conference on Historical Linguistics*, Montréal, août 2007. A paraître.

Monsonego S., Graff J., Derniame O., Henin M. (1989), La Lemmatisation assistée par ordinateur de textes de Moyen Français, 1 : Méthode générale. Nancy, Université de Nancy II.

Morphalou (2004), « Morphalou : un lexique morphologique ouvert du Français ». http://www.atilf.fr/morphalou/

Prévost S., Heiden S., Dupuis F. (2000), « Catégorisation d'un corpus hétérogène de français médiéval ». *Actes du colloque 'JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles'*, p. 477-487, Lausanne, 2000.

Schmid H. (1994), "Probabilistic Part-of-Speech Tagging Using Decision Trees". *Proceedings of International Conference on New Methods in Language Processing*, Manchester, England, p. 44-49, 1994. September 1994.

Kunstmann P., Stein A. (2006), « Le Nouveau Corpus d'Amsterdam », *Actes de l'atelier de Lauterbad*, Lauterbad, Allemagne, 23-26 février 2006.

Kunstmann P., Gerner H., Souvay G. (2007), « *DÉCT : Dictionnaire Électronique de Chrétien de Troyes »*, CILPR 2007 Congrès International de Linguistique et de Philologie Romane, Innsbruck, Autriche, 3-8 septembre 2007.
http://www.atilf.fr/dect/

Smith J. C. « Middle French : When ? What ? Why ? », in *Language Sciences* 24 (2002) 423-445.

Souvay G. (1986), « Analyse de textes de Moyen-Français », rapport de DEA, Centre de Recherche en Informatique de Nancy, Université de Nancy I.

Souvay G. (2004), « Vers un Dictionnaire électronique du Moyen Français », *Actes du Colloque Euralex 2004, European Association for Lexicography congress*, vol. 2 p.671-678. Lorient, France, 6-10 juillet 2004

Souvay G. (2007), « LGeRM : un outil d'aide à lemmatisation du français médiéval », *18th International Conference on Historical Linguistics ICHL 2007*, Université du Québec À Montréal. Canada. 6-11 août 2007.

Tobler A., Lommatzch E., Tobler-Lommatzsch : *Altfranzösiches Wörterbuch*, 1925.