Vers un Dictionnaire électronique du Moyen Français.

Gilles Souvay
ATILF-CNRS/Université Nancy 2
44 avenue de la libération
BP 30687
54063 NANCY CEDEX
FRANCE
gilles.souvay@atilf.fr

Résumé
L'équipe du Moyen Français du laboratoire ATILF (ex INaLF Nancy) a entrepris depuis 1990, la réalisation d'un Dictionnaire du Moyen Français (DMF). La première étape du projet a conduit à la réalisation de 13 lexiques réunis dans une base de données unique appelée Base de Lexiques du Moyen Français (BLMF). Elle est constituée de 69 000 articles pour 26 500 lemmes et a une taille de 87 millions de caractères. Elle est en accès libre sur l'internet depuis début 2003, à l'adresse http://www.atilf.fr/blmf.
Les données structurées au format XML sont exploitées par le moteur de recherche Stella (développé initialement pour le Trésor de la Langue Française informatisé et la base textuelle Frantext). L'utilisateur dispose ainsi d'un outil performant par la richesse dans l'expression de ses requêtes et par la rapidité des réponses.
Le contenu de cette base lexicale a ensuite été utilisé pour mettre au point un lemmatiseur capable de reconnaître 132 500 formes graphiques différentes ou d'interpréter à l'aide de règles morphologiques des formes inconnues.

1. Introduction

L'équipe du Moyen Français du laboratoire ATILF (Analyse et Traitement Informatique de la Langue Française) rédige des lexiques d'auteurs et des lexiques de genres, étape préalable à la réalisation d'un dictionnaire électronique évolutif du moyen français. Le moyen français est la langue utilisée dans la période 1350-1500.

2. Dictionnaire du Moyen Français

2.1. Projet initial

2.1.1. Les origines. Les lexiques constituent un ensemble de matériaux dans le but de réaliser le Dictionnaire du Moyen Français (DMF). Le DMF est un projet initié en 1957 à Strasbourg au « Colloque de lexicologie et lexicographie française et romane ». Le démarrage effectif n'a lieu qu'en 1990 à la fin de la rédaction du Trésor de la Langue Française (TLF). Il bénéficie de l'expérience du TLF : pour les rédacteurs (issus du TLF) et pour la méthodologie (une base textuelle source d'exemples pour rédiger les articles).

2.1.2. La rédaction. Les rédacteurs sont des membres de l'équipe du moyen français ou des collaborateurs extérieurs au laboratoire. Les lexiques sont saisis sous forme électronique avec un traitement de texte par des secrétaires scientifiques. Ensuite, une synthèse est réalisée par des rédacteurs du DMF, les lexiques sont alors considérés comme des réservoirs d'exemples

classés et interprétés pour le dictionnaire. Chaque lexique fait l'objet d'une publication papier indépendante (Gerner, 2003).

2.1.3. Les matériaux de rédaction. Les articles sont rédigés à partir de quatre types de matériaux :

- une base de textes à saisie intégrale dont le but est de refléter au mieux les différentes expressions écrites du moyen français par le choix des textes : romans en prose, en vers ; théâtre, religieux, profane ; œuvres poétiques ; traités ; chartes ; documents. Elle comporte 218 textes. Cette base est connue à l'extérieur de l'ATILF sous le nom de Frantext Moyen Français http://atilf.atilf.fr/dmf.htm ;
- une base de textes à saisie partielle qui complète la première en apportant les faits nouveaux concernant le vocabulaire qui n'aurait pas été rencontré dans la première base. Elle comporte 460 textes ;
- un glossaire des glossaires : une liste cumulative des entrées de 120 glossaires d'éditions critiques lemmatisées d'après Tobler-Lommatzsch, ou à défaut Godefroy ou le Complément du Godefroy ;
- des dossiers de mots : 35 000 dossiers non informatisés, réservoir d'exemples, de remarques contenant en outre le dépouillement de revues ou colloques, des études...

2.2. Réorientation du projet

2.2.1. Les raisons. En 1998 un bilan sur l'état d'avancement du projet a été effectué. À cette date, un premier volume du DMF avait été édité, couvrant la tranche de lettres A-AH, la tranche AI-AS était rédigée, 13 lexiques étaient achevés ou en voie d'achèvement (pour une trentaine effectivement lancés) dont 4 publiés. La valeur scientifique du travail a été reconnue, mais la durée du projet estimée à plus de 20 ans avant son achèvement a conduit à repenser le projet en 2001.

2.2.2. *Les nouvelles orientations.* De nouvelles orientations ont été données au projet. Les critères retenus ont été les suivants :

- donner une visibilité plus grande au projet,
- tenir compte des évolutions de l'informatique dans les outils et les techniques,
- donner une plus grande responsabilité aux rédacteurs.

Il a donc été décidé :

- que les lexiques devenaient autonomes et étaient regroupés en un seul produit,
- que les rédacteurs rédigeaient eux-mêmes leur lexique,
- que le traitement de texte était abandonné pour un outil de saisie balisée SGML/XML,
- que le produit était mis à la disposition de la communauté scientifique sous la forme d'une base de données en ligne en accès libre et d'un cédérom.

Le projet initial est appelé DMF^0 ; trois nouvelles étapes sont prévus DMF^1, DMF^2 et DMF^3.

2.2.3 L'étape DMF^1. L'étape DMF^1 consiste à publier sous forme électronique les 13 lexiques achevés. Les lexiques rédigés avec le traitement de texte sont rétroconvertis au format XML. Les entrées des lexiques sont lemmatisées pour regrouper les articles traitant du même lemme, les rédacteurs n'ayant pas forcément choisi la même graphie ; par exemple sous le lemme *âge* présent dans 13 lexiques, 9 ont choisi *age*, 3 *aage* et 1 seul *eage*. Chaque

lemme est pourvu de l'étymon sous lequel il est traité dans le FEW (Wartburg, 1922-2002) ; à défaut, dans la mesure du possible, un étymon est proposé.
La Base de lexiques du Moyen Français (BLMF) est le résultat de cette première étape. Elle est visible depuis janvier 2003 sur le site du laboratoire ATILF. Elle traite environ 26 000 lemmes.
2.2.4. L'étape DMF^2. La base est complétée à l'aide de nouveaux lexiques : 12 lexiques qui n'étaient pas achevés lors de l'étape 1 et un nouveau lexique appelé lexique de mots complémentaires. Le lexique complémentaire traite tous les lemmes qui ne sont pas étudiés dans les autres lexiques. Il est réalisé par l'ensemble des rédacteurs qui se répartissent le travail par tranche alphabétique.
Le DMF^2 traitera environ 60 000 lemmes et devrait être opérationnel avant 2006. La plus grosse partie du travail de rédaction est dors et déjà réalisée.
2.2.5. L'étape DMF^3. L'étape trois portera sur deux aspects : synthèse et mots grammaticaux. Il est en effet difficile de consulter un article qui a été traité dans plusieurs lexiques. Un travail de synthèse s'avère nécessaire pour faciliter la lecture des différents sens d'un lemme. Le travail portera sur 800 mots lourds traités dans 11 lexiques et plus soit 3,2 % de la nomenclature de la BLMF.
Les mots grammaticaux n'ont pas été étudiés. Quelques cas isolés seulement sont à l'heure actuelle présents dans la base. Cette lacune sera comblée à cette étape du projet.
La réalisation du DMF^3 devrait commencer en 2004 pour se terminer en 2007. La nomenclature ne devrait pas considérablement augmenter et devrait rester autour de 60 000 lemmes. La communauté scientifique disposera alors du DMFe : Dictionnaire du Moyen Français électronique. Le e minuscule signifie électronique, mais pourrait aussi vouloir dire évolutif. Du fait de l'utilisation de la norme XML, la correction des articles erronés et l'ajout de nouvelles données devraient être grandement facilités.

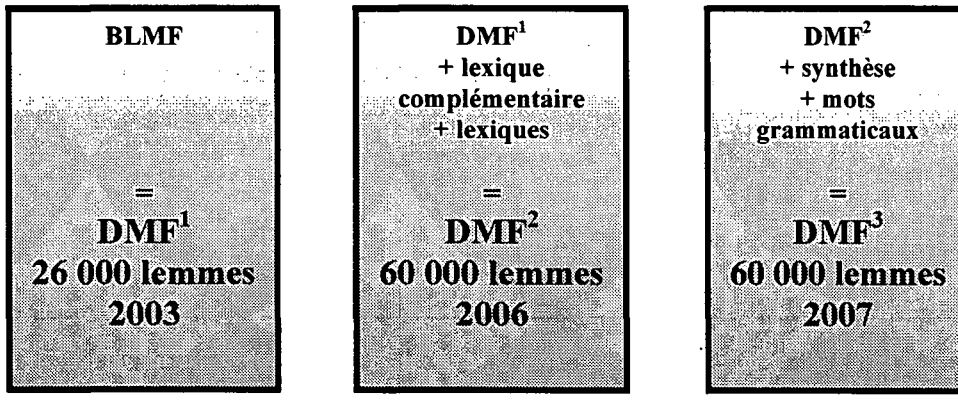

Figure 1 : Les étapes du nouveau DMF

3. Base de Lexiques

3.1. Saisie des lexiques

3.1.1. La structure des lexiques. Un lexique du moyen français est une collection d'articles au format XML. Chaque information pertinente est marquée à l'aide d'une balise spécifique et respecte une DTD (Document Type Definition : définition des balises utilisées et de leur enchaînement logique). Il y a deux types d'articles : les articles de renvoi et les articles standards.

Un *article de renvoi* a une structure simple : il est composé d'une vedette et d'une cible de renvoi.

Un *article standard* est composé d'une vedette, d'un code grammatical, d'un lemme, d'un rubrique de référence à des dictionnaires, d'une suite de paragraphes et d'une remarque globale à l'article qui est facultative. La balise dictionnaire comporte des références aux dictionnaires Tobler-Lommatzsch, Godefroy, DEAF, FEW et Trésor de la Langue Française. Un paragraphe est composé d'une partie métatexte (numérotation, indicateur, condition d'emploi, définition) et d'une série d'exemples (texte de l'exemple avec l'occurrence du lemme et référence bibliographique) et d'une remarque locale au paragraphe.

<ART> <VED>**FAUSSER** </VED> <CODE>**, verbe** </CODE> <LEM>**fausser** </LEM>
<DICT>[<TLGD>T-L, GD : <LEMME>*fausser* </LEMME> </TLGD> <GDC> ; GDC :
<LEMME>*falser* </LEMME> </GDC> <FEW.CONNU>; FEW <VOLUME>III, </VOLUME> <PAGE>393 :
</PAGE> <ETYM>*falsus* </ETYM> </FEW.CONNU> <TLF> ; TLF <VOLUME>VIII, </VOLUME> <PAGE>690a :
</PAGE> </TLF>] </DICT>
<P> <DISC> <IND>Empl. trans. </IND> <SYNT>*Fausser qqc.* </SYNT> </DISC> </P>
<P> <DISC> <NUM>**A.** - </NUM> <DEF>"Falsifier, contrefaire" </DEF> </DISC> <EXE> : <TEXTE>...se il avoit en cest accord aucunes paroles obscures, doubles ou autres par quoy il y <OCC>*faulsist* </OCC> déclaracion, les dites parties veulent et se assentent que elles soient interprétées, déclarées et amendées à l'entencion de maistre Jehan Canard </TEXTE> <REF> (Cartul. Laval B., t.2, 1384, 313) </REF>. </EXE> </P>
<P> <DISC> <NUM>**B.** - </NUM> <DEF>"Manquer à ses engagements ; trahir" </DEF> </DISC> <EXE> :
<TEXTE>...le dit suppliant la trouva en un certain heritaige qui fu au dit feu Guillaume Climent, son mary, et là corrocié et esmeu de chaude cole, et ayant souvenance et argu de ce que long temps paravant par fors induction avoit tant fait que la dicte femme avoit <OCC>*faussé* </OCC> son mariage envers le dit suppliant, comme il sceut depuis, dont il fut en voye de la tuer dès lors, soubz umbre de ce que elle cuilloit du fruit ou dit heritaige </TEXTE> <REF> (Doc. Poitou G., t.6, 1399, 345) </REF>. </EXE> </P> </ART>

Figure 2 : Un exemple d'article avec son balisage

3.1.2. L'outil de saisie. Le rédacteur dispose sur son poste de travail d'un outil appelé éditeur de texte balisé. Il lui permet de construire ses articles en enchaînant les balises tout en respectant les contraintes fixées par la DTD. Le rédacteur n'a pas à s'occuper de la mise en forme de son document. Les caractères gras, italiques, les ponctuations séparatrices... sont placées automatiquement par une feuille de style.

3.1.3. Les outils d'aide à la saisie. Le rédacteur construit à l'aide de l'éditeur un document correct par rapport à la DTD. Mais il n'a aucune assurance en ce qui concerne les informations contenues à l'intérieur même des balises. C'est pourquoi, il dispose d'un

auxiliaire précieux, l'outil de contrôle de la saisie, qui lui indique les erreurs qu'il a commises. Exemples de vérifications :

- valeur prise dans une liste prédéfinie (code grammatical, entrées du FEW…),
- règles de ponctuations correctes (usage de la virgule, guillemets ouvrants et fermant…),
- numérotation correcte des paragraphes…

L'outil de contrôle de la saisie est accessible depuis la toile. Il contient toute la documentation utile à la saisie.

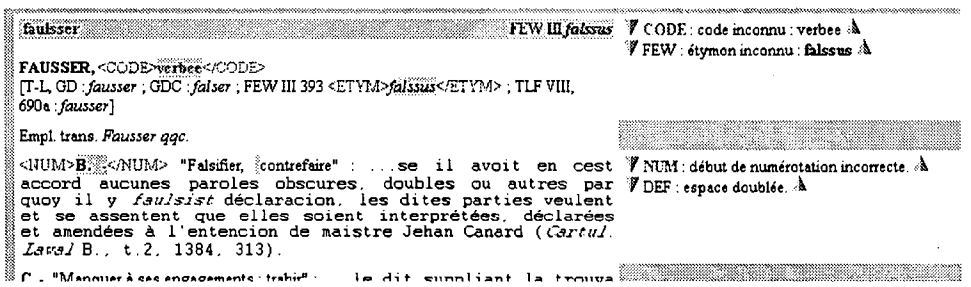

Figure 3 : Un exemple de contrôle

D'autres outils sont à la disposition du rédacteur. Ils permettent par exemple d'intégrer par simple copier/coller des extraits des bases textuelles, de parcourir le FEW, de consulter les références bibliographiques et de les importer…

3.1.4. Les données externes. Les données externes sont des informations supplémentaires qui ne sont pas gérées par le rédacteur. Elles complètent un article : l'étymon qui est rattaché directement au lemme et l'extension des références bibliographiques. Elles sont ajoutés automatiquement au moment de la construction de la base.

3.2. Mise en ligne des lexiques

3.2.1. Les caractéristiques de la base. Les lexiques sont accessibles sur internet par le biais d'une base sur la toile. La BLMF est en accès libre sur le site du laboratoire ATILF à l'adresse http://www.atilf.fr/blmf. Début mars 2004 elle est composée de :

- 13 lexiques,
- 26 351 lemmes,
- 34 551 vedettes,
- 69 090 articles,
- 70 726 graphies d'occurrence,
- 185 600 exemples,
- 192 920 occurrences,
- pour un total de 87 millions de caractères (balisage inclu).

À titre d'exemple pour le lemme *âge* :

- 3 vedettes : 9 *age*, 3 *aage*, 1 *eage*
- 13 articles (le lemme est traité dans les 13 lexiques)
- 74 exemples
- 75 occurrences (il y a deux occurrences dans un des exemples)

- 11 graphies d'occurrences différentes : 41 aage, 12 eage, 5 aages, 5 age, 3 aäge, 2 aaige, 2 eages, 2 acge, 1 aaiges, 1 aige, 1 eaige

3.2.2. Le moteur de recherche. Le moteur de recherche est le programme permettant d'interroger le texte balisé. Son choix est déterminant dans l'exploitation des données. Nous avons choisi le moteur Stella développé au laboratoire ATILF (Dendien, 2002). Le même moteur est utilisé pour le Trésor de la Langue Française informatisé (TLFi : http://www.atilf.fr/tlfi) et la base de données textuelle Frantext (http://www.atilf.fr/frantext). Stella est aussi une boite à outils facilitant les développements et offrant une interface avec la toile (Bernard et al., 2001).

Les principales qualités du moteur de recherche sont :

- sa fiabilité éprouvée à travers le TLFi et Frantext,
- le temps de réponse réduit en regard de la quantité de données à exploiter,
- la richesse dans l'expression des requêtes : recherche de graphie simple expressions régulières, choix dans une liste...

3.2.3. Interface. L'interface de la BLMF, quoique reprenant la boite à outils Stella, a fait l'objet de développements spécifiques aux lexiques et aux besoins des utilisateurs. Elle n'est pas définitivement figée et reste ouverte à des modifications qui pourraient être demandées par les membres de la communauté scientifique utilisant cette ressource.

Une effort a été porté sur la prise en main rapide de la base, sans avoir à connaître au départ la structuration des données ou encore la syntaxe d'une expression régulière... Nous avons aussi joué au maximum sur les possibilités de liens qu'offre un logiciel de navigation.

Ainsi deux grandes catégories d'expression d'une recherche sont mises en œuvre dans la BLMF : la recherche à partir des menus et la recherche par hyperlien dans le corps des articles.

3.2.4. La recherche à partir des menus. Lorsque l'utilisateur entre dans la base, il dispose d'un menu "Recherche dans la base". Trois types de recherche sont proposés.

La *recherche sur les entrées* porte sur le lemme, la vedette, l'étymon (pour la recherche dans une famille de mots), sur une graphie lemmatisée (occurrence) ou sur les lemmes du Tobler-Lommatzsch. Elle se déroule en deux phases. Dans un premier temps, il faut saisir une chaîne de caractères, la position de cette chaîne dans l'entrée (en début, à l'intérieur, en fin ou égalité stricte) et éventuellement indiquer si la recherche est sensible à la casse ou aux diacritiques. Après validation, la liste des entrées correspondant aux critères s'affiche, et l'utilisateur peut choisir alors d'afficher le ou les articles qui l'intéresse, ou d'affiner son filtre.

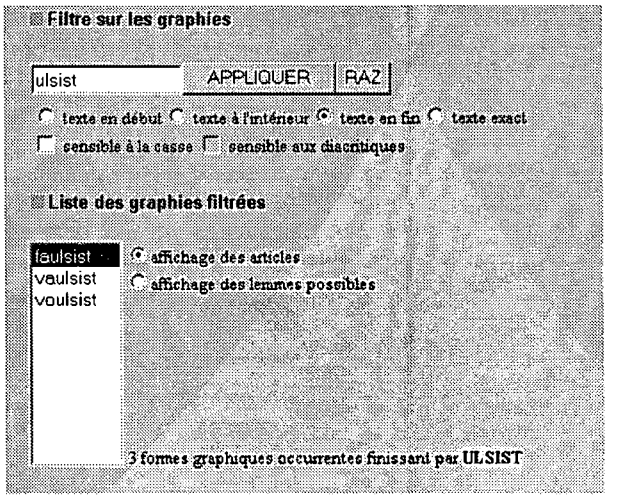

Figure 4 : Un exemple de recherche sur les graphies

La *recherche en plein texte* permet de trouver un mot sans tenir compte du balisage. On peut rechercher non seulement le texte exact, mais aussi un nom, un adjectif ou un verbe à fléchir. Attention, il s'agit pour l'instant d'une flexion moderne.

Enfin, la *recherche avancée* repose sur la richesse de Stella dans la formulation des requêtes. Elle nécessite une bonne connaissance de la structure des lexiques et des concepts de base de Stella. On peut combiner plusieurs critères, rechercher des suites de mots (expressions Stella), utiliser des expressions régulières, des listes déroulantes…

3.2.5. La recherche à partir du corps de l'article. Lorsque un article est affiché, l'utilisateur a la possibilité de cliquer sur des zones ou des mots qui le composent, plutôt que de revenir aux menus :

- en cliquant sur l'étymon, il a accès aux mots de la même famille,
- en cliquant sur le lemme, il a accès à la liste des graphies attachées au lemme,
- en sélectionnant une référence bibliographique, l'extension complété de la référence apparaît dans une nouvelle fenêtre,
- en cliquant sur un mot de l'article, il déclenche la lemmatisation du mot
- etc…

3.2.6. Affichage de l'article. L'article (ou les articles) résultat de la recherche est affiché dans un style aussi varié que le serait une édition papier en jouant sur les polices, le gras et l'italique, les couleurs... L'article est découpé en trois zones :

- en début d'article une bande grisée contenant le lemme et l'étymon ;
- en fin d'article une seconde bande grisée contenant le nom du lexique, et l'auteur de l'article ;
- entre les deux bandes, on retrouve les principaux composants de l'article : vedette et code sur la première ligne, référence aux dictionnaires sur la deuxième et ensuite une série de paragraphe ; à l'intérieur des paragraphes les occurrences du lemme sont marquées en bleu italique.

FAUSSER FEW III *falsus*

FAUSSER, verbe

[T-L, GD : *fausser* ; GDC : *falser* ; FEW III, 393 : *falsus* ; TLF VIII, 690a]

Empl. trans. *Fausser qqc.*

A. - "Falsifier, contrefaire" : ...se il avoit en cest accord aucunes paroles obscures, doubles ou autres par quoy il y *faulsist* declaracion, les dites parties veulent et se assentent que elles soient interprétées, déclarées et amendées à l'entencion de maistre Jehan Canard (*Cartul. Laval* B., t.2, 1384, 313).

B. - "Manquer à ses engagements ; trahir" : ...le dit suppliant la trouva en un certain heritaige qui fu au dit feu Guillaume Climent, son mary, et là corrocié et esmeu de chaude cole, et ayant souvenance et argu de ce que long temps paravant par fors induction avoit tant fait que la dicte femme avoit *faussé* son mariage envers le dit suppliant, comme il sceut depuis, dont il fut en voye de la tuer dès lors, soubz umbre de ce que elle cuilloit du fruit ou dit heritaige (*Doc. Poitou* G., t.6, 1399, 345).

Chartes et Coutumes Edmonde Papin *balisage fin*

Figure 5 : Un exemple d'article

4. Lemmatiseur

4.1. Présentation

Les lexiques ont été construits essentiellement à partir de deux bases textuelles. Que se passerait-il si on projetait leur contenu (plus exactement les graphies lemmatisées) sur ces textes ? De cette idée est né le projet de développer un lemmatiseur spécifique au Moyen Français. C'est une langue difficile à traiter automatiquement car elle présente une très grande variété de formes fléchies pour un lemme. Par exemple pour le lemme agneau, on peut trouver les formes agnel, agnels, aignel, aignels, agneaus, aigneaux, agneauls, aigneaulz, aingnel... Il n'y a pas d'outil connu dans la communauté essentiellement à cause de l'absence d'un dictionnaire de référence. Au laboratoire ATILF, nous disposons d'une première version d'un dictionnaire et des graphies lemmatisées en fonction des entrées de ce dictionnaire.

4.2 Algorithme et base de connaissance

4.2.1 L'algorithme. Le premier objectif que nous nous sommes fixé est de fournir la bonne analyse du mot à lemmatiser parmi une liste de possibilités. L'algorithme mis en œuvre consiste à rechercher le mot, hors contexte, parmi l'ensemble des graphies lemmatisées dont nous disposons. Si le mot n'est pas une graphie connue, le programme applique des règles morphologiques pour le ramener à une forme connue. Il existe des mécanisme de rejets pour éviter d'appliquer des règles de flexion verbales sur des lemmes non verbaux, ou pour éviter un trop grand nombre d'application de règles.

En résumé, le lemmatiseur est un système à base de règles. La base de connaissance sur lequel il s'appuie est constitué d'une base de graphies lemmatisées et d'un ensemble de règles morphologiques.

4.2.2 Les graphies lemmatisées. En mars 2004, le lemmatiseur s'appuie sur près de 132 500 graphies :

- les occurrences dans la BLMF (environ 82 600),
- les occurrences dans les lexiques en cours de rédaction, les lexiques non exploités, le DMF^0 (environ 47 000),
- une liste de mots outils constituée manuellement (environ 800),
- des graphies additionnelles (ajout manuel de graphies qui ne correspondent pas à des mots outils) (environ 400),
- une liste de noms propres (ajout manuel ou extraits de Frantext) (environ 1 700).

4.2.2 Les règles morphologiques. La définition des règles morphologiques s'appuie au départ sur des travaux réalisés en 1986 par l'Équipe de l'Unité de Recherche sur le Français Ancien, Unité de Recherche Linguistique N 10. Université Nancy 2 (Souvay, 1986). Elles sont de la forme : *si* condition *alors* action *finsi*. La condition est facultative, la règle s'applique alors systématiquement. Une action consiste à transformer une suite de lettres en une autre suite.

Exemples de règles :

si précédé de [A,E,O,U] *alors* LL se transforme en L *finsi*
PP se transforme en P
si finale *alors* OIT se transforme en ER *finsi*
si précédé de voyelle et suivi de voyelle *alors* U se transforme en V *finsi*

Exemple d'analyse : étudions le cas de l'analyse de la graphie *cuilloit*. Trois hypothèses ont été engendrées par le lemmatiseur.
La première hypothèse a transformé la graphie *cuilloit* en *cuiller* en imposant qu'il s'agisse d'une flexion verbale. La base de graphies contient la forme *cuiller* mais avec un code grammatical substantif. Elle est donc rejetée.
La deuxième hypothèse a transformé la graphie *cuilloit* en *cueillir* en en imposant qu'il s'agisse d'une flexion verbale. La base de graphies contient la forme *cueillir* avec un code grammatical verbal. L'hypothèse est donc acceptée.
La troisième hypothèse a permis après application de 3 règles morphologiques de passer de la forme *cuillot* à *chevilloit* qui est dans la base de graphies associée au lemme *cheviller*. Un nouveau mécanisme est alors mis en œuvre. Le lemmatiseur s'aperçoit que 3 règles morphologiques ont été appliquées. Il considère alors que c'est trop comparé à l'unique règle qui a conduit au succès de la seconde hypothèse. Il rejette alors l'hypothèse *cheviller*.
Seules les hypothèses ayant conduit à un succès temporaire sont présentées. En interne le lemmatiseur travaille sur beaucoup plus de formes. En particulier dans le cas de cuilloit, il a engendré 143 graphies : cuylloit, chuilloit, seuilloyt, chouilloit ...

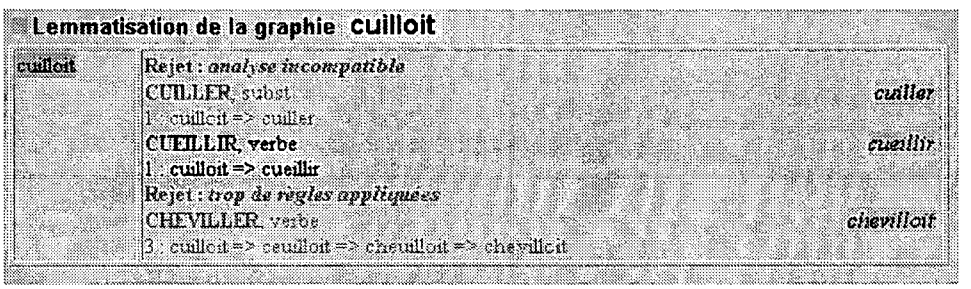

Figure 6 : Analyse de la graphie *cuilloit*

De la centaine de règles initiales, on arrive aujourd'hui à près de 400 règles concernant uniquement des règles de variations graphiques (doublement d'une lettre, omission d'une lettre, forme pluriel). En ce qui concerne la flexion verbale, leur nombre n'est pas déterminé avec précision aujourd'hui, mais il y en a plusieurs milliers.

4.3 Mise en œuvre

Un premier prototype du lemmatiseur est accessible en ligne depuis la base textuelle Frantext Moyen Français et depuis la BLMF. Dans Frantext Moyen Français ou dans un article de la BLMF, un clic sur un mot provoque l'ouverture d'une nouvelle fenêtre et l'affichage du lemme avec accès aux articles le concernant.
La BLMF dispose d'une entrée Lemmatisation qui permet de saisir une portion de texte et de lancer sa lemmatisation.

4.4 Évaluation

Le lemmatiseur n'a pour l'instant fait l'objet d'aucune évaluation. Il semble être néanmoins assez performant et arrive dans la grande majorité des cas à fournir une analyse pertinente de la graphie. Il peut être pris en défaut sur les formes conjuguées qui n'ont pas encore été inventoriées exhaustivement.

4.5 Projet de base textuelle lemmatisée

Un projet de lemmatisation de textes de Moyen Français est en cours de définition. L'utilisation du lemmatiseur combiné avec d'autres outils, devrait permettre de constituer un corpus Frantext Moyen Français Lemmatisé selon les lemmes de la BLMF. Le lemmatiseur a pour défaut de ne pas lever l'ambiguïté entre une forme verbale et un substantif (analyse hors contexte), par exemple s'il rencontre la forme *devoir*, il ne peut dire s'il s'agit d'un substantif ou d'un verbe. Néanmoins il peut constituer une première étape pour un traitement ultérieur par un outil plus performant.

5. Conclusion et perspectives

Une première version du Dictionnaire de Moyen Français électronique est désormais accessible librement sur l'internet à l'adresse http://www.atilf.fr/blmf. Il s'agit d'un dictionnaire entièrement informatisé de la saisie à la publication. Il est balisé au format XML ce qui facilitera son évolution future. Il intègre un moteur de recherche puissant, Stella, qui lui permet d'accéder facilement et rapidement au cœur des articles où se situe l'information pertinente. Il sera complété d'ici 2007 pour passer de 26 000 lemmes aujourd'hui à près de 60 000.

Le dictionnaire a donné naissance à un outil de lemmatisation. Cette outil n'est encore qu'un prototype, qui devra faire ses preuves. Il se révèle relativement performant dans la reconnaissance des formes mais pêche dans la désambiguïsation des hypothèses concurrentes. Il devrait être utilisé, sans doute en association avec d'autres outils, pour développer des corpus de textes lemmatisés.

Références

ATILF/Équipe "Moyen français et français préclassique", 2003/2004. *Base des lexiques du moyen français*, site internet (http://www.atilf.fr/blmf).

Bernard P., Dendien J., Pierrel J.-M., Souvay G., Tucsnak Z., 2001. Les ressources informatisées de l'ATILF pour l'étude du français: dictionnaires, encyclopédie, bases textuelles et logiciels d'exploitation. *Actes du séminaire Corpus et ressources de l'ILF*, Paris, juin 2001, « *Institut de Linguistique Française : Corpus, ressources méthodes et outils* », p. 61-76

Colloque Lexicologie et lexicographie françaises et romanes, Strasbourg, Octobre 1957, Publication du CNRS, Paris: Éds. CNRS

Dendien, J., 2002. STELLA et ses fonctionnalités, Congrès international de Rouen, L'édition électronique en littérature et dictionnairique: évaluation et bilan (17-22 juin 2002).

Gerner H., Souvay G. 2003. Base de données textuelles et base de données lexicales en moyen français sur le site du laboratoire ATILF, in: P. Kunstmann / F. Martineau / D. Forget (éds): Ancien et moyen français sur le Web : enjeux méthodologiques et analyse du discours (Voix savantes; 20). Ottawa: Les Éd. David, 147-161

Souvay, G. 1986. Analyse de Textes de Moyen-Français. Rapport de DEA, Université de Nancy.

Wartburg W., 1922-2002. FEW, Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes, 25 vol., Bonn, Klopp/Berlin, Teubner/Bâle, Zbinden