Dictionnaires informatisés : les pratiques au laboratoire ATILF

Gilles Souvay

Dictionnaires informatisés : les pratiques au laboratoire ATILF

Gilles Souvay

ATILF (Analyse et Traitement Informatique de la Langue Française), CNRS (Centre National de la Recherche Scientifique) et UL (Université de Lorraine).

gilles.souvay@atilf.fr

1. Présentation

L'ATILF (Analyse et Traitement Informatique de la Langue Française) est un laboratoire de recherche en sciences du langage. Cette unité mixte de recherche (UMR 7118) a deux tutelles : le Centre national de la recherche scientifique (CNRS) et l'Université de Lorraine (UL). Il est situé dans le nord-est de la France à Nancy. Le site du laboratoire est à l'adresse <www.atilf.fr>.

Le laboratoire existe sous ce nom depuis 1981. Il est né dans les années 60 dans le but de réaliser le *Trésor de la Langue Française*, un dictionnaire de la langue française du 19e et 20e siècle.

L'ATILF s'appuie sur une double compétence forte en linguistique et en informatique, qui lui permet de mettre à disposition des chercheurs de nombreux outils en ligne, dont plusieurs dictionnaires et autres ressources de référence dans ses domaines de recherche.

Cet article présentera une sélection de ressources lexicographiques produites à l'ATILF avec quelques aspects techniques de leur réalisation et leur environnement informatique.

2. Ressources informatisées

Le ressources informatisées de l'ATILF se déclinent selon trois grands domaines : la lexicographie en ligne, les bases de données textuelles et les outils pour le TAL (Traitement Automatique des Langues).

2.1. Lexicographie en ligne

2.1 1. Des éléments

Les dictionnaires en ligne sont une des grandes spécialités de l'ATILF qui a été un des laboratoires pionniers dans le domaine avec le TLFi, la version informatisée du TLF, ouverte en 2002. Par la suite d'autres dictionnaires informatisés ont été réalisés à l'ATILF pour les états anciens du français (langue médiévale) et dans le domaine de l'étymologie française et romane.

Ces dictionnaires sont des références non seulement dans leur contenu scientifique mais aussi dans la méthodologie utilisée. En ce qui concerne le contenu scientifique des linguistes spécialistes du domaine au niveau national et international ont contribué à la rédaction des articles. En ce qui concerne la méthodologie, recours à des outils informatiques s'appuyant sur une structuration des données en XML et utilisation d'une plate-forme de diffusion interactive

Sous les termes *dictionnaire informatisé* ou *dictionnaire en ligne* se cache un ensemble de réalisations très différentes selon leur implémentation informatique, en terme d'accès au contenu, d'ouverture vers d'autres ressources, en terme d'évolutivité...

On peut déjà distinguer les dictionnaires en *mode image* par opposition aux dictionnaires en *mode texte*. Pour un dictionnaire en mode image, on prend une photo de chacune des pages, on fait une liste des entrées et enfin on associe à chaque entrée, la première page de l'article. C'est le reflet d'un dictionnaire physique existant, souvent créé avant le développement de l'informatique. Au laboratoire ATILF un exemple de ce type de dictionnaire est la version électronique du FEW (Französisches Etymologisches Wörterbuch) <https://apps.atilf.fr/lecteurFEW/>. La communauté des médiévistes utilise le *Dictionnaire de Godefroy* et sa version image <micmap.org/dicfro/introduction/dictionnaire-godefroy>. Ce type de dictionnaire permet de mettre rapidement en accès un ouvrage, avec des coûts en temps de développement réduits, mais en contrepartie les fonctions d'accès au contenu du dictionnaire sont limitées. Pour un dictionnaire en mode texte, on rencontre deux cas de figure. Il y a d'un côté les dictionnaires existants avant le développement de l'informatique, il faut les ressaisir, ou les numériser avec des logiciels de reconnaissance de caractères. De l'autre côté, il y a les dictionnaires bénéficiant d'une saisie informatique native. Dans les deux cas, il faudra repérer chaque entrée comme on le fait pour un dictionnaire en mode image. Mais pour les dictionnaires en mode texte, on peut complexifier l'informatisation en

repérant la structure et les différents éléments de l'article : entrée, code grammatical, définition, domaine, exemple… On parle dans ce cas de figure de balisage des informations et on utilise le langage XML (Extensible Markup Language).
En résumé on a d'un côté les dictionnaires en mode image et de l'autre les dictionnaires en mode texte, structurés ou non.

La recherche dans un dictionnaire informatisé est très liée à son implémentation et à la structuration ou non des articles. Pour un dictionnaire traditionnel papier, un seul mode d'accès : l'entrée. Quand on recherche le sens d'un mot, on doit trouver son entrée, en général le singulier pour un substantif, le masculin singulier pour un adjectif, l'infinitif pour un verbe... Pour un dictionnaire informatisé, on peut faire de la même façon laisser l'utilisateur taper l'entrée mais on peut aller plus loin. Tout d'abord pour la recherche sur l'entrée, on peut gérer la flexion et la variation graphique des mots de manière assistée. On recherche *pensée*, le dictionnaire va non seulement proposer d'aller consulter le substantif *pensée* (la fleur ou l'activé psychique) mais aussi le verbe *penser*, il a supposé que *pensée* était le participe passé féminin de ce verbe. Pour résoudre informatiquement ce problème, on utilise souvent des listes de mots (des lexiques morphologiques : mot, lemme et flexion du mot). Pour les dictionnaires informatisés en mode texte, on peut rechercher dans le corps de l'article, en plein texte, ou alors si le dictionnaire a été balisé, un mot ou toutes les flexions d'un mot dans un élément de l'article. Exemple de recherche : quelles sont les entrées qui contiennent le mot *aimer* dans leur définition. On imagine bien que dans un dictionnaire papier, ou un dictionnaire informatisé non balisé on va devoir y passer beaucoup de temps.

Une autre force d'un dictionnaire informatisé, c'est qu'on peut combiner facilement plusieurs critères de recherche. Exemple de recherche : trouver tous les verbes du domaine de la botanique.

Un dictionnaire en ligne sur internet permet de faire des renvois à d'autres ressources informatisées. Des liens vers d'autres dictionnaires : un dictionnaire spécialisé dans les domaines techniques, dans les régionalismes, un dictionnaire concurrent, un dictionnaire étymologique, un dictionnaire en diachronie. Des liens vers d'autre type de ressources : morphologie, synonymie, concordance dans des corpus textuels… C'est ce que propose le

portail lexical du CNRTL (Centre National de Ressources Textuelles et Lexicales), basé à l'ATILF, autour du TLF <www.cnrtl.fr/portail>.

Une chose que le responsable scientifique du dictionnaire informatisé ne doit pas négliger, c'est l'évolutivité du contenu. Le dictionnaire est-il figé, ou peut-il lui apporter des corrections, ajouter de nouveaux articles ? Peut-il exporter ses données dans un format standard ? Pour faire cette correction, est-il entièrement autonome, ou doit-il attendre la disponibilité du responsable informatique de son projet ? C'est un aspect qu'il ne faut pas négliger, nombre de projets sont freinés voir arrêtés par un informaticien indisponible…

Les articles d'un dictionnaire sont illustrés avec des exemples d'utilisation. Va-t-il s'agir d'exemples créés de toute pièce par le rédacteur de l'article, ou va-t-il aller les chercher dans la littérature, dans un corpus textuel dédié ? En plus de la constitution du corpus, qui peut s'avérer être une tâche lourde, il faudra prévoir une composante bibliographique.

Différents dictionnaires existent à l'ATILF, présentant ou non plusieurs des caractéristiques mentionnées plus haut. En ce qui concerne les dictionnaires développés au sein de l'équipe linguistique historiques, ils sont gérés avec une plate-forme possédant les caractéristiques les plus avancées : un dictionnaire en mode texte finement balisé, un moteur de recherche permettant de parcourir la structure, une gestion détaillée de la bibliographie, des corpus textuels associés (Souvay ; Renders 2014).

2.1.2. Le Trésor de la Langue Française

Le TLF est un dictionnaire du française des XIX^{e} et XX^{e} siècles. La rédaction du TLF est terminée depuis 1994 et la plupart des contributeurs ont quitté le laboratoire. Il n'a pas vocation à être mis à jour. Cette ressource, qui ne fait pas l'objet d'une veille lexicographique, est donc close « en l'état ». Il comporte 100 000 mots avec leur histoire, 270 000 définitions, 430 000 exemples issus de la littérature française, ce qui représente environ 350 millions de caractères. Il est publié en version papier en 16 volumes.

La version informatisée du *Trésor de la Langue Française* est appelée TLFi. Il possède une structuration fine des données, les différents éléments de l'article sont balisés (définition, conditions d'emploi, domaines, exemples, source des exemples... Les données du TLF ont été entièrement ressaisies et relue. Des programmes automatiques ont posés les balises qui

ont été ensuite vérifiées. Le TLFi est paru initialement en 2012 dans une version cédérom désormais obsolète avec le développement grand public d'internet et la possibilité d'offrir un accès libre au dictionnaire.

Deux sites utilisent les données du TLFi, chacun ayant sa propre approche. Tout d'abord la version historique à l'adresse <www.atilf.fr/tlfi>. Elle n'a pas évolué depuis son ouverture, les interfaces sont anciennes, pas toujours ergonomiques, mais c'est la plus élaborée pour les recherches en particulier sur la structure, la combinaison de critères… La seconde version via le portail lexical du CNRTL <www.cnrtl.fr/portail> est plus récente du point de vue technologie utilisée, mais tout aussi figée depuis son ouverture, et ne propose que la recherche sur les entrées, avec néanmoins prise en compte de la flexion. Le plus de cette version comme indiqué plus haut, est sa connexion avec d'autres ressources informatisées : morphologie, diachronie, diatopie, synonymie, corpus textuel…

2.1.3. D'autres dictionnaires à l'ATILF

Deux grandes orientations sont présentes à l'ATILF, les dictionnaires étymologiques et les dictionnaires en diachronie/synchronie.

En ce qui concerne les dictionnaires étymologiques, l'ATILF héberge le FEW (Französisches Etymologisches Wörterbuch). Le FEW écrit par le philologue suisse Walther von Wartburg est le principal dictionnaire étymologique de référence des langues gallo-romanes. La création de l'édition originale du FEW, s'est amorcée en 1922 et achevée en 1967. Il est accessible en mode image à partir des entrées. Un mode texte avec un balisage fin est en cours de réalisation (Renders 2011).

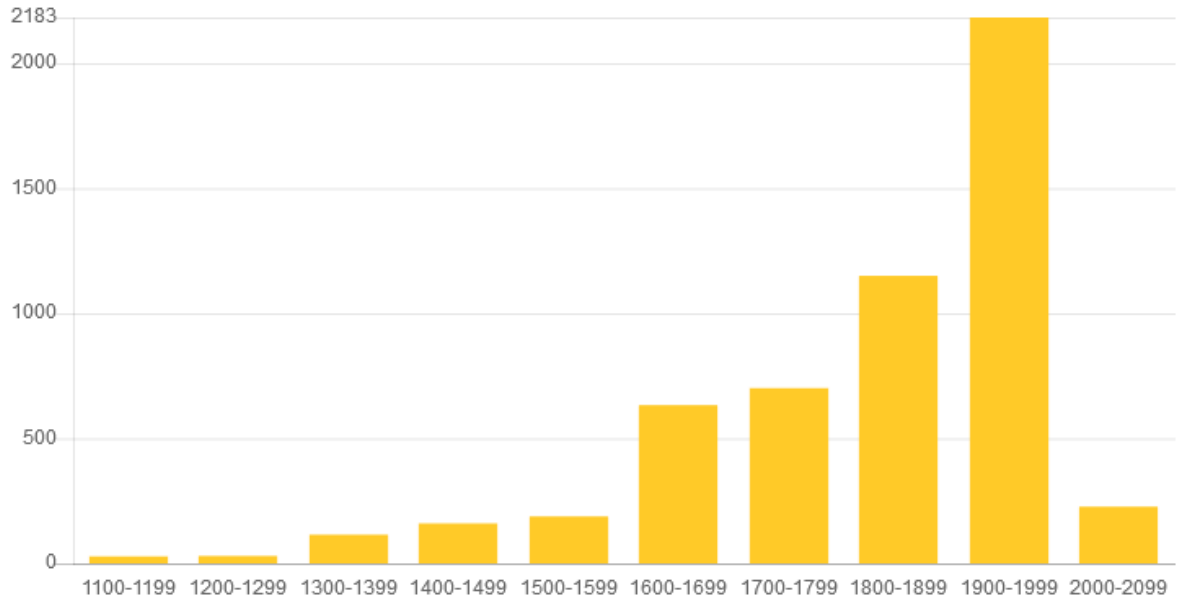
Un second dictionnaire de taille plus réduite consiste à mettre à jour la rubrique étymologique du TLFi. Il s'agit du programme de recherche TLF-Étym <www.atilf.fr/tlf-etym/>. Ce dictionnaire contient à actuellement 526 notices étymologiques. Le dictionnaire est structuré en XML et en plus de l'entrée, il peut être interrogé sur une sélection de critères. Il est implémenté avec la plate-forme ISIS, qui permet au responsable du projet de gérer le dictionnaire, et le mettre à jour aussi souvent que nécessaire en déposant la version XML des articles qui sont alors immédiatement disponibles (Souvay ; Renders 2014).

Un troisième dictionnaire qui couvre l'étymologie pan-romane a été développé dans le cadre de projets ANR (Agence nationale de la recherche)/DFG(Deutsche Forschungsgemeinschaft). Il s'agit du DÉRom (Dictionnaire Étymologique Roman). Les articles sont balisés en XML, on peut interroger les entrées et des champs prédéfinis. Comme le dictionnaire précédent, il est implémenté avec la plate-forme ISIS.

En ce qui concerne les dictionnaires en diachronie du français, deux principaux ouvrages sont en ligne à l'ATILF. Pour la période Ancien Français, le *Dictionnaire Électronique de Chrétien de Troyes* (DÉCT) <www.atilf.fr/dect> constitue à la fois un lexique complet de cet écrivain du XII^{e} siècle et une base textuelle qui permet de lire ou d'interroger les transcriptions de ses cinq romans (*Érec*, *Cligès*, *Lancelot* ou le *Chevalier à la Charrette*, *Yvain* ou le *Chevalier au Lion*, *Perceval* ou le *Conte du Graal*). Le second dictionnaire en diachronie concerne la période du moyen français (1330-1500) avec le *Dictionnaire du Moyen Français* <www.atilf.fr/dmf> et couvre l'ensemble de la langue. Ils sont construits sur le même modèle : des schémas de définition du balisage proches, un balisage fin des structures des articles ; des recherches possibles dans la structure ; un corpus textuel ayant permis de rédiger les articles ; de nombreux liens vers d'autres ressources lexicales et textuelles. Ils ont été saisis nativement en XML et sont diffusés avec la plate-forme ISIS.

2.2. Corpus textuel : Frantext

La base de données textuelle Frantext <www.frantext.fr> est une des références dans le domaine des corpus textuels. Initialement créée dans les années 1970 afin de fournir des exemples pour le TLF puis pour le DMF, elle a poursuivi son développement et constitue désormais une ressource linguistique à part entière. Elle est constituée de textes littéraires, techniques, scientifiques pour toutes les périodes de français. En juin 2019 il comportait 5 415 références. Elle comporte des textes sous droits, d'où un accès contrôlé avec abonnement.



Répartition chronologique par siècle des textes dans Frantext

2.3. Ressources pour le TAL

L'ATILF distribue différentes ressources utiles pour les outils du traitement automatique des langues et en particulier des lexiques morphologiques pour toutes périodes du français : Morphalou pour la période moderne <www.ortolang.fr/market/lexicons/morphalou> et LGeRM pour le français médiéval et le français du XVIIe <www.ortolang.fr/market/lexicons/lgerm>. Morphalou est utilisée par le CNRTL pour gérer la flexion des entrées et présenter la morphologie des mots. LGeRM est utilisé par le DMF et le DÉCT pour gérer la flexion et la variation graphique des entrées.

L'ATILF diffuse aussi les fichiers paramètres pour l'étiqueteur ayant permis d'annoter le corpus Frantext. *Modèle Talismane pour textes littéraires en français moderne* <www.ortolang.fr/market/tools/talismane-frantext-modern>

D'autres ressources sont disponibles, elles sont listées sur le site de l'ATILF <www.atilf.fr>.

3. Dictionnaire du Moyen Français

Nous allons maintenant présenter le *Dictionnaire du Moyen Français* et les aspects techniques spécifiques qui ont dû être résolus pour qu'il soit facilement consultable. C'est un

des dictionnaires électroniques de référence pour la langue médiévale, et une grande partie de la communauté des médiévistes l'utilise.

3.1. Un dictionnaire informatisé à tous les niveaux

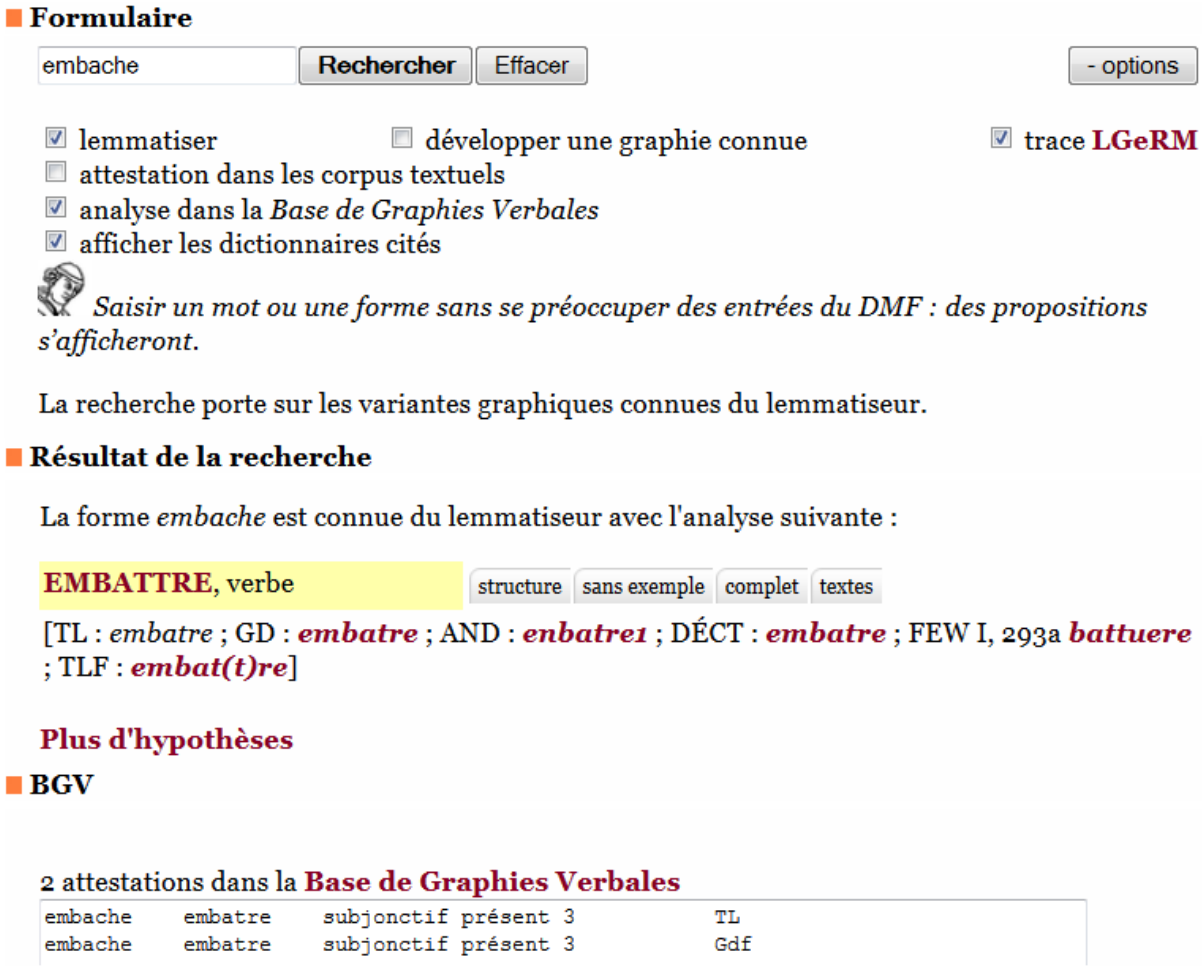
La principale difficulté en diachronie est de pouvoir trouver le mot que l'on recherche dans le dictionnaire que l'on consulte. Pour le français contemporain, on maitrise en général la langue, on connait sa morphologie, il y a peu de variations graphiques, on arrive facilement à trouver l'entrée sous laquelle le mot est traité sans vraiment besoin d'assistance. En ce qui concerne la langue médiévale, on maitrise beaucoup moins la morphologie et on est surtout confronté à la variation graphique des mots due à l'absence de norme graphique et aux marquages régionaux encore très présents dans les textes, surtout pour les plus anciens.

Le problème pour trouver l'entrée correspondant au mot que l'on cherche se pose aussi bien au spécialiste de la langue médiévale qu'à l'étudiant débutant dans l'étude de cette période du français. Quelle entrée (quel lemme) a été retenue pour le dictionnaire ? Pour une forme comme *destroict* on peut supposer qu'il faut regarder sous le substantif *détroit*, pour *ameroyent*, le verbe *aimer*, c'est d'autant plus facile que les mots existent encore dans la langue contemporaine. En ce qui concerne *acormens* on est en face d'un mot ayant disparu du français contemporain, en connaissant un peu la morphologie on aurait envie de regarder sous *acourment*, *accourrement* ou encore *acourement*… on a le problème des lettres doublées ou pas. Face à *polra*, il faut connaitre la morphologie verbale pour deviner le verbe *pouvoir*, mais face à *aulter* on est un peu démuni avec cette variante régulière du lemme *autel*.

Le choix de la graphie du lemme est aussi un problème pour les rédacteurs du dictionnaire. Pour le mot *agneau* moderne, une entrée *agnel* serait tout à fait légitime. L'équipe de rédaction du DMF a choisis initialement de moderniser les graphies des mots anciens pour faciliter la consultation du dictionnaire et essayer de garder une cohérence dans une famille, surtout au niveau des mots disparus. Malheureusement la construction du dictionnaire étape par étape, la durée du projet, l'impossibilité parfois de trancher, les changements de points de vue au cours du temps n'ont pas permis d'être toujours cohérent.

La solution retenue pour la consultation informatique du DMF est algorithmique. L'utilisateur tape la forme telle qu'il l'a rencontrée dans le document, sans se soucier de savoir où le mot est rangé dans le dictionnaire. Le mot est lemmatisé à la volée et le DMF propose d'aller consulter une ou plusieurs entrées du dictionnaire.

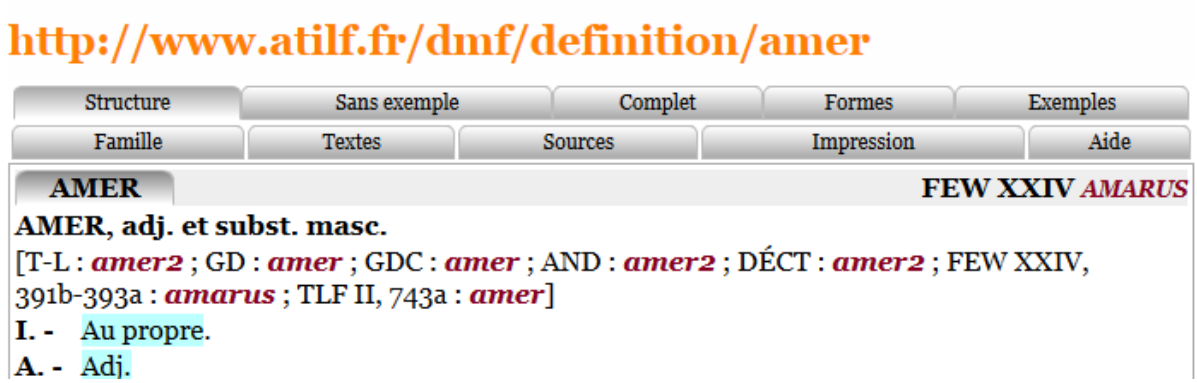
Exemple : chercher le mot *embache* dans le DMF :



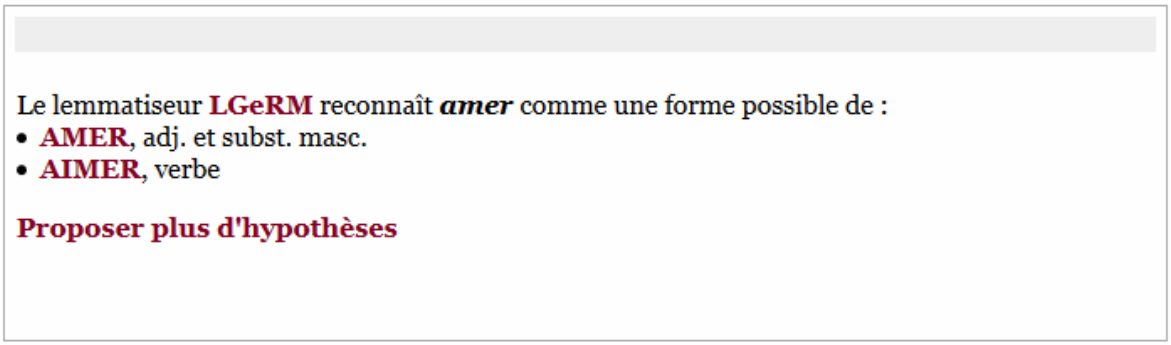
L'informatique est au cœur de l'élaboration du dictionnaire La conservation des exemples et leur sélection se fait à travers des bases de données textuelles. La rédaction des articles se fait en XML avec un éditeur de texte balise, on n'utilise pas de traitement de texte qui ne

permet pas facilement de marquer la structure des articles. La consultation du dictionnaire est assistée par un lemmatiseur.

Le DMF est en accès libre à l'adresse <www.atilf.fr/dmf>. La dernière version est datée de 2015. Il contient 65 720 entrées, 470 125 exemples, soit environ 200 millions de caractères. Cela représente, si on l'imprimait 19 900 pages, soit l'équivalent de 15 volumes du TLF. Les points forts du dictionnaire sont sa bonne couverture de toute l'étendue de la langue médiévale, sa gestion performante de la variation graphique et de la morphologie médiévale, les nombreux liens qui le connecte aux différentes autres ressources médiévales de la communauté (Tobler-Lommatzsch <http://as-bwc-tl.spdns.org/tl/>, Anglo Norman Dictionary <http://www.anglo-norman.net/>, The Online Froissart < https://www.dhi.ac.uk/onlinefroissart/>…), et la possibilité à la communauté de connecter le dictionnaire à ses propres ressources à l'aide d'adresse pérennes :
<www.atilf.fr/dmf/definition/amer> pour afficher l'article *amer*.



<www.atilf.fr/dmf/morphologie/amer> pour analyser le mot *amer*.



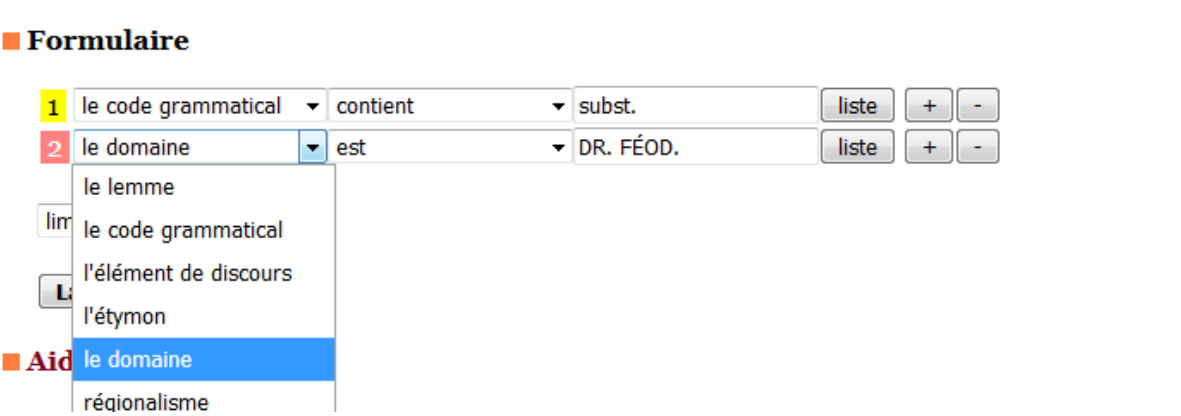
Derrière les articles du DMF se cache une structure XML qui peut être interrogée via les interfaces du dictionnaire. La structuration ne suit pas les recommandations de la TEI (Text Encoding Initiative). Une première raison est que la première version du DMF est antérieure aux recommandations. Néanmoins s'il y avait besoin, le XML serait facilement transformable. Une seconde raison est que le DMF utilise sa propre plate-forme et n'a pas besoin d'être dans ces normes. Et enfin, il s'avère que le schéma de saisie avait pour but initialement de faciliter la saisie des lexicographes qui sont passés d'une saisie faite avec un traitement de texte par une secrétaire, à une obligation de saisir eux-mêmes les articles prêts à monter. Le DMF est un des tous premiers dictionnaires saisie nativement en XML.

Le balisage de l'article FIEF :

FIEF, **subst. masc.** **fief**
[T-L, GDC : ***fief*** ; FEW XV-2, 117a : ****fehu*** ; TLF VIII, 843b : ***fief***]
A. - DR. FÉOD.
1. "Domaine noble relevant d'un suzerain que celui-ci concède en tenure à un vassal (en dehors de toute rente) en contrepartie de l'hommage et du service requis" : Del ***fief*** l'empereor estes a tort saisis (Garin Lorr. M., c.1330-1400, 486). ...deux cenz livres de rente, les quelles le dit conte avoit vendues au dit cardinal, assises, selon la coustume du païs, en la chastelerie de Syvray, avec toute justice, haute, moienne et basse en ***fiez*** et arrerefiez, et à touz autres droiz, quiex qu'il feussent, excepté tant seulement ressort et souveraineté (Doc. Poitou G., t.2, 1335, 125). ...je suis si courcié que sommes si meschant Que n'ay terre, ***fief***, ne ung chastel vaillant (Ren. Gennes D.B., c.1350-1400, 89). Vo ***fief*** en croisteray d'une riche contree. (Renaut Mont. B.N. V., c.1350-1400, 366). Tant avoit richesse et puissance, Terres, ***fiez*** [var. ***fies***], honneur et avoir Que trop estoit de tant avoir. (MACH., C. ami, 1357, 29). Il donnoit ***fiez***, joiaus et terre, Or, argent ; riens ne retenoit Fors l'onneur ; ad ce se tenoit, Et il en avoit plus que nuls. Des bons fu li mieudres tenus. (MACH., C. ami, 1357, 103). ...mon nepveu (...) Met si a non chaloir le sien Que de ses ***fiez*** et heritages Ne li chaut (Mir. chan., c.1361, 141). Et certes, selon la Loy civile, et selon raison, le vassal qui ne fait le service que il doit a

ART VED **FIEF** /VED CODE, **subst. masc.** /CODE LEM **fief** /LEM
DICT [TLGDC T-L, GDC : LEMME ***fief*** /LEMME /TLGDC FEW.CONNU ; FEW VOLUME XV-2, /VOLUME PAGE 117a : /PAGE ETYM ****fehu*** /ETYM /FEW.CONNU TLF ; TLF VOLUME VIII, /VOLUME PAGE 843b : /PAGE LEMME ***fief*** /LEMME /TLF] /DICT
P DISC NUM A. - /NUM DOM DR. FÉOD. /DOM /DISC /P
P DISC NUM 1. /NUM DEF "Domaine noble relevant d'un suzerain que celui-ci concède en tenure à un vassal (en dehors de toute rente) en contrepartie de l'hommage et du service requis" /DEF /DISC EXE :
TEXTE Del OCC ***fief*** /OCC l'empereor estes a tort saisis /TEXTE REF (Garin Lorr. M., c.1330-1400, 486) /REF. /EXE EXE TEXTE ...deux cenz livres de rente, les quelles le dit conte avoit vendues au dit cardinal, assises, selon la coustume du païs, en la chastelerie de Syvray, avec toute justice, haute, moienne et basse en OCC ***fiez*** /OCC et arrerefiez, et à touz autres droiz, quiex qu'il feussent, excepté tant seulement ressort et souveraineté /TEXTE REF (Doc. Poitou G., t.2, 1335, 125) /REF. /EXE EXE TEXTE ...je suis si courcié que sommes si meschant Que n'ay terre, OCC ***fief*** /OCC, ne ung chastel vaillant /TEXTE REF (Ren. Gennes D.B., c.1350-1400, 89) /REF. /EXE EXE TEXTE Vo OCC ***fief*** /OCC en croisteray d'une riche

Exemple de requête avancée dans le DMF, rechercher les substantifs utilisés dans le domaine du droit féodal :

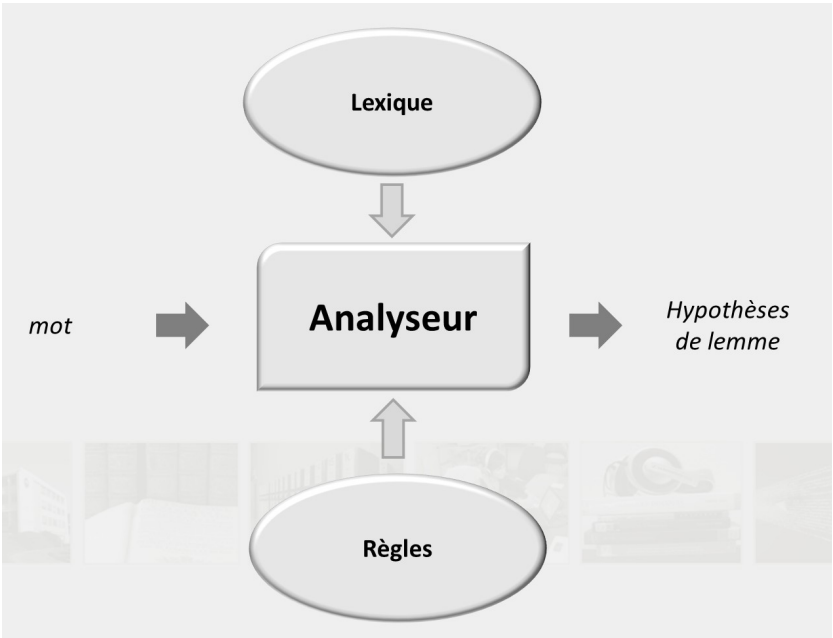


3.2. LGeRM : le point d'entrée du DMF

LGeRM est l'outil de lemmatisation qui permet de consulter le DMF. LGeRM est l'acronyme pour Lemmes, Graphies et Règles Morphologiques www.atilf.fr/LGeRM (Souvay ; Pierrel 2009).

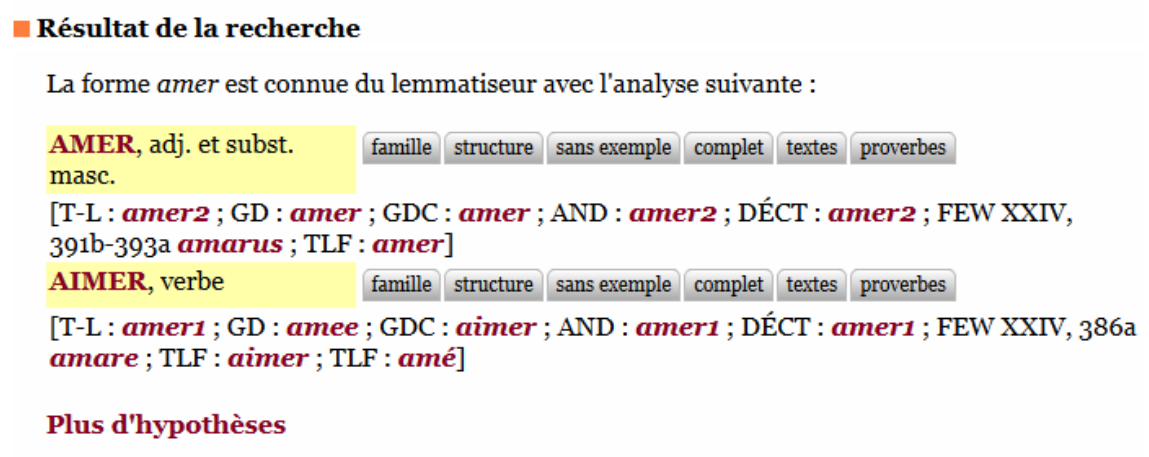
3.2.1. Architecture générale du lemmatiseur

On fournit un mot à un analyseur. Il utilise un lexique de formes connues et des règles morphologiques et de variations graphiques pour proposer des hypothèses de lemme pour le mot.

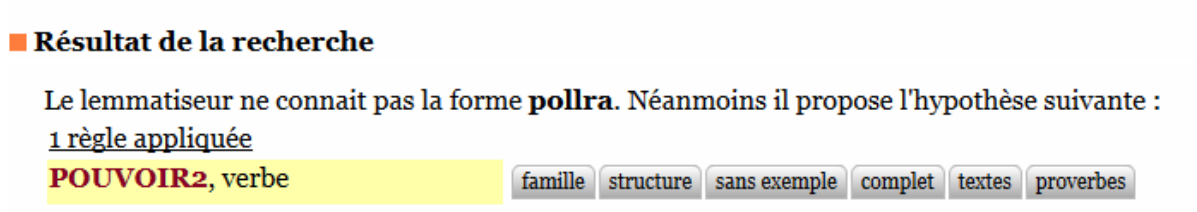


L'analyseur est un programme informatique. L'algorithme est plutôt simple. Si le mot est dans le lexique, il propose les analyses (les lemmes) connues pour ce mot. Si le mot n'est pas dans le lexique, il applique des règles de transformation du mot pour trouver un mot présent dans le lexique. Il faut un mécanisme d'arrêt du système pour éviter de boucler, pour éviter d'appliquer trop de règles et une stratégie de gestion des formes produites, en effet beaucoup de règles peuvent s'appliquer sur un mot. La couverture du lexique est suffisante pour en général proposer la bonne hypothèse à l'utilisateur du DMF. Néanmoins, si ce dernier n'est pas satisfait de la réponse, il peut demander au DMF de proposer de nouvelles hypothèses sur les mots connus. Cela corrige éventuellement les lacunes du lexique si le lemme est présent dans le dictionnaire.

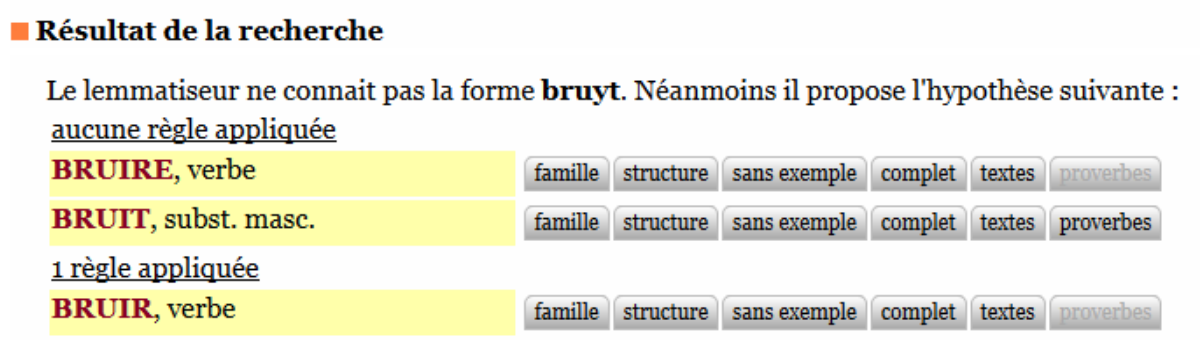
LGeRM connait deux lemmes pour la forme *amer.*



La forme *pollra* n'est pas dans le lexique, mais LGeRM propose le lemme pouvoir :



La forme *bruyt* est connue avec les lemmes bruit et bruire. En forçant LGeRM à appliquer les règles, il peut proposer en alternative le lemme bruir.



3.2.2. Le lexique

Le lexique LGeRM est une liste de triplets (forme, lemme, code grammatical). Exemples de triplets :

(amer, aimer, verbe)

(amer, amer, adj.)

(amera, aimer, verbe)

Les codes grammaticaux sont ceux des lemmes du DMF. Ils sont utilisés par les règles qui sont liées à la morphologie. Le lexique s'est construit et enrichi au fur et à mesure du développement du DMF. Les articles étant balises il était facile d'extraire une première liste de triplet. Par la suite des compléments ont été faits, très peu par procédures automatiques, mais plutôt en exploitant les formes présentes dans les corpus textuels, dans les différents textes traités par le lemmatiseur lors de collaborations formelles ou informelles avec des éditeurs de textes. Le lexique s'appuie donc sur des corpus textuels, plus spécifiquement Frantext, des formes réellement attestées dans les textes. En août 2019, le lexique comportait environ 975 500 entrées.

3.2.3. Les règles

Il existe deux grandes familles de règles : les règles portant sur la morphologie, flexion des mots et les règles portant sur la variation graphique. Un ensemble initial de 200 règles a été défini en s'appuyant sur les travaux réalisés dans le cadre d'un DEA *Analyse de textes de moyen-français* (Souvay 1986). Cet ensemble ne contenait aucune règle sur la morphologie verbale. Lors de l'ouverture du DMF, il a donc été nécessaire de compléter cet ensemble initial. Par la suite chaque mot demandé au DMF qui n'était pas correctement reconnu, chaque texte traité par le lemmatiseur, a permis d'ajouter de nouvelles règles. Actuellement

encore de nouvelles règles sont ajoutées, très souvent des cas particuliers de variantes spécifiques à un mot, à une graphie régionale ou plus spécifique à l'ancien français. En tout, il y a environ 6 500 règles, dont les trois quart portent sur la flexion verbale et sa variation.

La structure générale d'une règle est de la forme « si des conditions sont remplies alors on effectue une action » :

si *conditions* alors *action* finsi

Les règles peuvent s'appliquer sans condition. Les conditions portent sur le graphème du mot : l'initiale, la finale, les caractères qui entourent le graphème (précédé de, suivi de, liste de lettres, d'une consonne, d'une voyelle, sauf…). Les conditions peuvent porter sur le lemme résultat (son initiale, sa finale, une liste de lemmes…). Enfin une condition de succès ou non de la règle : si une règle peut s'appliquer sur un mot, on ne l'applique effectivement que s'il y a une solution, cela permet de réduire le champ d'exploration du lemmatiseur sur des règles qui s'appliqueraient trop souvent. En ce qui concerne les actions, on peut supprimer le graphème ou transformer le graphème en une autre suite de caractères.

Exemples de règle sans condition:

Y→I	*fayre*→ *faire,* FAIRE	modernisation
C→SS	mesfacent → *mesfassent*, MÉFAIRE	équivalence graphique
OUN→ON	mount → mont	variante anglo-normande

Pour ce qui traite de la flexion, la première approche serait d'essayer de retrouver l'infinitif du mot. Il existe en effet des règles de cette nature. Mais compte tenue de la variation sur la base, il nous a semblé plus pertinent de créer des règles de transformation de la personne, du genre, du nombre…

si *en finale* alors ES→EF finsi	nes→nef, NEF	flexion des lemmes en –EF
si *en finale* alors ERA→ER finsi	amera→amer, AIMER	infinitif pour IF3S (indicatif futur 3ème personne du singulier)
si *en finale* alors RONT→RA finsi	menront→menra, MANER	passage de IF3P à IF3S
si *en finale* et précédé de [D, T, V]alors ERAI→RAI finsi	ponderai→pondrai, PONDRE	variation de la flexion

Les règles sont utilisées pour pallier les lacunes du lexique. Il semble impossible, tant la variation est grande, d'établir un lexique exhaustif de toutes les formes médiévales. Toute la variation graphique est contenue dans les règles. À titre d'exemple, si l'on voulait décrire la flexion et variation du lemme connaissance, on obtiendrait une expression régulière ressemblant à :
[c|k|q][o|oi|e|oei][n|nn|gn|ngn][oi|ai|i|ioi|e|oe][s|ss|sc|sç|ç|c][i]?[en|an|ã|ẽ][s|ss|c|sc|ç|ch][e][sz]?

Il y a 55 formes identifiées dans les corpus médiévaux de l'ATILF : *cognescence cognissance cognissanche cognoeissance cognoiscences cognoisçance conaisanche congnioissance congnoessance connissanche conoissances cougnoissance*…

L'objectif de LGeRM est d'identifier toutes la variation et flexion d'un lemme connu. Il n'est pas un outil figé. Il s'enrichit en permanence de nouvelles formes, de nouvelles règles, voire de nouveaux lemmes.

L'outil est disponible sur demande pour des tests, un travail est en cours pour le rendre distribuable plus largement. Il existe une version en ligne limitée à quelques mots accessible depuis le DMF. Enfin il utilisable via une plate-forme de lemmatisation qui permet de traiter les éditions nouvelles ou anciennes numérisées que la communauté continue de produire. Il s'avère être un bon relecteur pour les erreurs d'océrisation ou de saisie, un outil pratique pour établir le glossaire d'une édition, et un bon moyen de diffuser un texte lemmatisé et étiqueté.

Le lexique est distribué librement sous licence Creative Common. Deux versions sont disponibles, une version focalisée sur le français médiéval et une seconde focalisée sur le français du XVII^{e} <www.ortolang.fr/market/lexicons/lgerm>.

4. Conclusions

L'ATILF a une expertise dans la création de ressources et dans leur mise à disposition sur internet et plus spécialement en lexicographie informatisée. Elle diffuse des dictionnaires traitant de la langue moderne ou sur l'histoire de la langue, diachronie et étymologie. Elle a produit des outils de référence qu'elle ouvre et diffuse largement pour les linguistes en lexicographie, corpus textuels ou traitement automatique de la langue.

Sites internet cités

Site du laboratoire ATILF : <www.atilf.fr>

Französisches Etymologisches Wörterbuch en ligne : <apps.atilf.fr/lecteurFEW/>

Portail lexical du Centre National de Ressources Textuelles et Lexicales : <www.cnrtl.fr/portail>

Trésor de la langue française informatisé : <www.atilf.fr/tlfi>

Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle, Frédéric Godefroy, 1880-1895 : <micmap.org/dicfro/introduction/dictionnaire-godefroy>

Le programme de recherche TLF-Étym : <www.atilf.fr/tlf-etym/>

Dictionnaire Électronique de Chrétien de Troyes : <www.atilf.fr/dect>

Dictionnaire du Moyen Français (1330-1500) : <www.atilf.fr/dmf>

Base de données textuelle Frantext : <www.frantext.fr>

Lexique morphologique Morphalou : <www.ortolang.fr/market/lexicons/morphalou>

Lexiques morphologiques LGeRM : <www.ortolang.fr/market/lexicons/lgerm>

Modèle Talismane pour textes littéraires en français moderne : <www.ortolang.fr/market/tools/talismane-frantext-moderne>

Dictionnaire électronique en ligne Tobler-Lommatzsch : <as-bwc-tl.spdns.org/tl>

The Anglo-Norman On-Line Hub : <http://www.anglo-norman.net>

The online Froissart : <https://www.dhi.ac.uk/onlinefroissart>

Plate-forme de lemmatisation LGeRM : <www.atilf.fr/LGeRM>

Bibliographie

RENDERS, Pascale (2011) *Modélisation d'un discours étymologique. Prolégomènes à l'informatisation du Französisches Etymologisches Wörterbuch*. Thèse de Nancy-Université et Université de Liège.

SOUVAY, Gilles (1986) *Analyse de textes de moyen-français*. Rapport de Diplôme d'Études Approfondies. Centre de Recherche en Informatique de Nancy. Université de Nancy I.

SOUVAY, Gilles ; PIERREL, Jean-Marie (2009) *LGeRM : lemmatisation des mots en moyen français.* Traitement Automatique des Langues, volume 50, numéro 2. <halshs.archives-ouvertes.fr/halshs-00396452/document>

SOUVAY, Gilles ; RENDERS, Pascale (2014) *Traitement informatique du DÉRom*. in Buchi, Eva; Schweickard, Wolfgang (Eds.) Dictionnaire Étymologique Roman (DÉRom). Genèse, méthodes, résultats.