De l'outil-glossaire à la plate-forme de lemmatisation (2007-2021)

Sylvie Bazin et Gilles Souvay

À partir de 2007, la rédaction du *Dictionnaire du Moyen Français* a pu se poursuivre grâce au travail de longue haleine de Robert Martin et des principaux collaborateurs qu'il lui restait, mais également grâce au soutien du laboratoire et à notre investissement sur place. Cinq nouvelles versions du *Dictionnaire du Moyen Français* ont vu le jour depuis lors (2009, 2010, 2012, 2015, 2020). Robert Martin, dans son témoignage, en indique les principales caractéristiques. Il est important de souligner également le travail souterrain de montage des articles et l'évolution de l'interface lors des différentes versions du *Dictionnaire*. Mais, au-delà de la continuité de ce grand projet, élaboré antérieurement, que nous avons toujours défendu, le laboratoire attendait l'émergence d'un nouveau projet scientifique.

Créé par G. Souvay pour répondre à la difficulté de la variation graphique, tant dans l'élaboration que dans la consultation des articles du *Dictionnaire*, le lemmatiseur LGeRM a été constamment repris et amélioré au fil de ces années. Il a élargi son champ d'intervention grâce à toute une série de collaborations scientifiques et a participé à la production de nouvelles ressources : il est le cœur désormais d'une plate-forme de lemmatisation accessible sur la page d'accueil du DMF ou par un lien direct : **http://www.atilf.fr/dmf**

En premier lieu, grâce au projet de coopération international, initié par Hiltrud Gerner, « Le *Dictionnaire de moyen français* et autres dictionnaires de langues vernaculaires médiévales : principes, méthodologie, pratique, problèmes et solutions » avec les Universités d'Edinburgh, de Liverpool et Sheffield (2007-2011) un premier outil d'aide à l'édition et à la construction de glossaires, a pu être construit. Il s'agissait d'épauler deux gros projets d'édition en ligne :

https://www.dhi.ac.uk/onlinefroissart/
http://www.pizan.lib.ed.ac.uk/

Dans le prolongement de cette collaboration, tous les ans depuis 2011, le texte médiéval au programme des agrégations de lettres et de grammaire est lemmatisé et traité avec les outils du DMF, ce qui permet d'offrir à la consultation une édition numérisée et d'interroger le texte par forme et par lemme, avec des liens vers les articles du DMF. Depuis 2021, l'interrogation peut également se faire par étiquette morphosyntaxique. Voici la liste de ces éditions lemmatisées : les *Poésies* de Charles d'Orléans, *Le roman de Tristan* de Béroul, *Le Roman de la Rose* de Guillaume de Lorris, *Le Couronnement de Louis*, *le Roman d'Eneas*, *Le Roman de Guillaume de Dole*, *Le Livre du Duc des vrais amants* de Christine de Pizan, *Yvain* de Chrétien de Troyes, *Les Lais* de Marie de France, *La Chanson d'Aspremont*, *Les Poésies* de François Villon, *La Mort du Roi Arthur*. Soit une grande variété de textes, écrits tant en AF qu'en MF et de genres différents (poésie, roman, chanson de geste)[1]. Pendant la période, d'autres éditions électroniques ont vu le jour. Elles sont liées à des collaborations originales, on trouvera la liste des liens sur l'onglet **Textes** de la page d'accueil du DMF. On peut ainsi signaler *Le Pèlerinage de Vie Humaine* de Béatrice Stumpf, le *Corpus des récits de voyage* de Capucine Herbert, ou encore le *Journal de Gouberville*.

L'outil utilisé conserve des liens étroits avec le DMF, mais il présente aussi des développements originaux. Créé pour la gestion de la variation dans le dictionnaire, il a élargi son champ d'intervention. Les textes médiévaux traités n'appartiennent pas tous à la période du moyen français, qui est la période de référence du *Dictionnaire du Moyen Français* ; des textes d'ancien français ont pu être lemmatisés avec des ajustements mineurs. L'outil glossaire a également connu des développements au-delà de la période de référence : même si les vérifications et les interventions manuelles sont plus importantes que pour le moyen français, LGeRM peut s'adapter aux textes antérieurs (ancien français) comme aux textes postérieurs (16e et 17e siècles). Pour ces derniers, la participation de l'ATILF au projet européen IMPACT[2] a été déterminante. En effet, pour chacun des pays partenaires, la collaboration entre bibliothèque nationale et laboratoire de linguistique historique avait pour but d'améliorer les logiciels d'océrisation utilisés sur les documents anciens. Dans ce cadre, l'ajout de formes

[1] Gilles Souvay est responsable pour chacun de ces textes du traitement automatique et ses adaptations, Sylvie Bazin de la relecture et de la correction de la lemmatisation automatique, à l'exception du texte de Chrétien de Troyes traité en collaboration avec Pierre Kunstmann et adossé non pas au DMF mais au DeCT (*Dictionnaire électronique de Chrétien de Troyes*) et des deux derniers, produits en collaboration avec Jean-Michel Jézéquel et pour le dernier, également avec Corinne Denoyelle, de l'Université Grenoble Alpes.

[2] *Improving Access to Text* : http://www.impact-project.eu

modernes à LGeRM grâce à une autre ressource de l'ATILF, MORPHALOU[3], a permis de construire un lexique capable de couvrir très largement la période choisie pour l'expérimentation (17e siècle). Le lexique moderne archaïsé a été projeté sur un corpus textuel issu de FRANTEXT et sur un corpus de seize textes ayant conservé leur graphie d'origine, numérisés dans le cadre du projet pour constituer la "vérité terrain". L'utilisation combinée des ressources a montré son efficacité dans le processus d'adaptation à une période intermédiaire entre le moyen français et la langue couverte par le TLF et MORPHALOU. Depuis, le lemmatiseur a été utilisé dans d'autres projets d'envergure, parce qu'il permet de traiter justement les textes des périodes anciennes de la langue : PRESTO, projet Phraséologie de Grenoble, *etc.* Ce faisant, et grâce à ces collaborations successives, le lemmatiseur n'est plus seulement un outil au service de la lexicographie, mais déploie ses possibilités dans le champ du Traitement automatique des langues. Désormais deux lexiques peuvent être distribuées, l'un pour la langue médiévale (AF et MF) et l'autre pour la période 16e -17e s.

Tous ces développements se greffent sur un projet initial, celui du DMF, qui a su prendre le tournant d'une lexicographie véritablement évolutive, en dépassant la construction lettre par lettre du dictionnaire au profit d'une construction globale par étape, mais aussi en faisant éclater les limites traditionnelles de la consultation, grâce au balisage des données et aux liens hypertextuels[4]. Au cœur du dispositif, le dictionnaire proprement dit demeure la source et la référence, à travers des versions datées qui sont archivées et une dernière version, directement disponible en ligne, qui reprend, corrige et enrichit la précédente. Cependant, en rappelant l'importance du lemmatiseur LGeRM et les développements qu'il a permis, nous évoquons un chantier de recherche encore en cours, qui progresse comme un chantier de fouille à partir de nouveaux sondages et de nouvelles campagnes liées à de nouveaux partenariats avec des éditeurs de textes anciens ou d'autres projets de recherche.

L'adaptation d'un tel outil à la langue et aux textes du moyen français, puis l'élargissement de son domaine d'intervention s'inscrivent bien dans le dynamisme et le réalisme d'une lexicographie évolutive qui n'abandonne pas les projets ambitieux, mais choisit de les faire avancer par étape, à la fois *dans* et *pour* une communauté de recherche, avec une ouverture vers un public élargi. LGeRM, accessible en ligne, documenté et illustré par des exemples, permet d'analyser des formes, de lemmatiser un extrait ou un texte complet, de construire un index lemmatisé, voire de réaliser un glossaire électronique de manière assistée. L'utilisateur, selon ses besoins, choisit le mode de collaboration, plus ou moins libre ou étroite. Ainsi notre ambition a-t-elle été de construire, sans toucher au noyau lexicographique – le *Dictionnaire du Moyen français* conçu et dirigé par Robert Martin – une plate-forme de lemmatisation, véritable espace en ligne de recherche au service des textes de moyen français, compris au sens large (et non plus au sens restreint des limites chronologiques 1330-1500). Les témoignages des chercheurs rencontrés montrent qu'ils ne s'arrêtent pas à ces bornes, que la consultation du DMF comme l'utilisation de la plate-forme de lemmatisation concernent également les spécialistes d'ancien français comme ceux du 16e s., sans parler aussi d'utilisateurs non spécialistes qui cherchent à comprendre les textes du passé. Le nombre de nos utilisateurs réguliers[5] et celles des collaborations effectives nous confortent dans l'idée de faire une œuvre utile. Comme le Moyen Âge aime à le dire, nous sommes des « enfants sur des épaules de géants », nous espérons grâce aux matériaux patiemment rassemblés et mis en forme par nos devanciers, et avec les convictions que nous nous sommes forgées, contribuer modestement à faire découvrir les textes anciens, à mieux les comprendre et à mieux les exploiter linguistiquement.

Sylvie Bazin et Gilles Souvay
(novembre 2021)

[3] Le lexique MORPHALOU est un lexique des formes fléchies du français, à large couverture (540.000 formes). Les données initiales proviennent du *TLFnome*, la nomenclature du *Trésor de la Langue Française*. Il est en accès libre à des fins de recherche et d'enseignement et sa mise à jour est assurée par l'ATILF.
[4] Voir la présentation de Robert Martin.
[5] Plusieurs centaines d'utilisateurs par jour. Quand le site est inaccessible pendant un certain temps, on reçoit de nombreux messages d'inquiétude, parce que les utilisateurs comptent désormais sur cet outil.