Informatisation du *Französisches Etymologisches Wörterbuch*

Depuis le 14 février 2014, le *Französisches Etymologisches Wörterbuch* (FEW) de Walther von Wartburg est consultable gratuitement en mode image sur internet, à l'adresse www.atilf.fr/lecteurFEW (voir également http://www.atilf.fr/few). Il s'agit de la première étape visible d'un projet dont l'objectif final est de rendre accessible la totalité du FEW en version électronique, accompagnée de fonctionnalités de recherche avancées et d'aides à la lecture. Nous décrivons ci-dessous la genèse du projet et l'état d'avancement de l'informatisation des 25 volumes, ainsi que les fonctionnalités prévues dans l'interface de consultation du futur FEW électronique.

1. Genèse du projet

On sait depuis longtemps (cf. Wooldridge 1990, 239) que l'informatisation du FEW résoudra une grande partie des difficultés d'accès à l'ouvrage. Parallèlement à la mise en chantier de projets portant sur certains aspects du dictionnaire (saisie du fichier onomasiologique, version électronique du *Complément* etc. : cf. Chauveau 2006), une étude de faisabilité portant sur l'informatisation des 25 volumes a été entamée en 2006, sous la forme d'une thèse de doctorat en cotutelle entre l'Université de Liège et l'Université de Nancy. Elle a conduit à la modélisation du discours lexicographique propre au FEW, à la formalisation de ce modèle au moyen d'un balisage XML et au développement d'un logiciel capable d'insérer les balises de façon totalement automatisée dans le texte du dictionnaire (Renders 2011). Le schéma XML a été pensé de façon à non seulement refléter les structures complexes de l'oeuvre, mais aussi répondre au maximum – dans les limites d'une automatisation – aux

doi: 10.2143/TD.87.0.3112860.

besoins des utilisateurs. Il identifie une quarantaine de types d'information, parmi lesquels les étymons-vedettes, les lexèmes galloromans et non-galloromans, les étiquettes géolinguistiques, les sigles bibliographiques, les auteurs d'articles, les étymons cachés ou encore les affixes.

La modélisation du FEW a pour particularité de rendre compte de deux visions complémentaires du FEW, que nous appelons d'une part sa *dimension thesaurus* (le FEW vu comme un ensemble d'unités lexicales) et d'autre part sa *dimension monographique* (le FEW vu comme un ensemble d'articles retraçant chacun l'histoire d'une famille lexicale). Ces deux visions correspondent à deux étapes généralement successives de consultation de l'ouvrage : d'abord la recherche d'une lexie particulière, voire de lexies partageant une ou plusieurs caractéristique(s), ensuite la lecture des articles où apparaissent cette ou ces lexie(s). La prise en compte de ces deux dimensions a une conséquence directe sur le balisage XML qui est inséré dans le discours fewien. Ce balisage est, en effet, tenu de réexpliciter un certain nombre d'informations qui, dans le discours lexicographique, sont implicites, soit parce qu'elles ont été citées précédemment et ne sont pas répétées (règle d'ellipse valable notamment dans l'infrastructure du FEW, cf. Büchi 1996, 77-78 ; 98), soit parce qu'elles sont déductibles de leur insertion dans un ensemble plus vaste (micro-, macro- et superstructure). Les étiquettes géolinguistiques, signifiants, signifiés et codes grammaticaux implicites sont par exemple systématiquement rétablis pour chaque unité lexicale, afin d'assurer l'exhaustivité d'une recherche en dimension thesaurus. En ce qui concerne la dimension monographique, ce sont par exemple les noms de rédacteurs, l'appartenance linguistique des étymons ou encore les marqueurs alphanumériques soutenant l'organisation microstructurelle des articles qui sont systématiquement réexplicités.

Le logiciel de balisage a été testé sur un corpus d'une centaine d'articles relevant notamment du projet ANR DETCOL (Développement et Exploitation Textuelle d'un Corpus d'Oeuvres linguistiques, voir http://ctlf.ens-lyon.fr/documents/de_anr_workshop-corpus.asp) et formant un échantillon

considéré comme assez varié pour valider l'étude. Les résultats très positifs (cf. Renders 2011, 292-302) ont donné le coup d'envoi au chantier d'informatisation du FEW.

2. Avancement de l'informatisation : un chantier en collaboration

L'informatisation proprement dite du FEW a débuté en 2012, sous la forme d'une collaboration entre l'Université de Liège et l'ATILF (CNRS & Université de Lorraine). La direction du projet est confiée depuis le 1er mai 2014 à Pascale Renders, chargée de recherches FNRS à l'Université de Liège, en étroite collaboration avec la direction du FEW. Ci-dessous sont mentionnées entre parenthèses les personnes directement impliquées dans chacune des composantes du projet.

Le processus d'informatisation suit concrètement trois étapes successives :

1. l'acquisition du texte brut des articles, accompagné d'un balisage typographique minimal (italiques, grasses, sauts de page et de colonne etc.) ;
2. l'insertion automatique, dans ce texte, du balisage XML complet ;
3. la mise en ligne des articles balisés, accompagnée d'une interface de consultation et d'un lien vers une version image du FEW.

La première étape s'effectue sous la forme d'une double saisie manuelle. Cette solution, plus efficace qu'une numérisation sujette à de nombreuses erreurs en raison, notamment, de la présence de nombreux caractères spéciaux, a été proposée et apportée par le Center for Digital Humanities de l'Université de Trèves (http://kompetenzzentrum.uni-trier.de). L'avancement de cette étape dépend des financements octroyés et ne suit pas nécessairement l'ordre de numérotation des volumes : le mois de juin 2014 a par exemple vu s'achever la saisie des volumes 16 et 17 de la partie germanique, ainsi que du volume 19 contenant les Orientalia. Ces trois volumes ont

d'abord été scannés à l'ATILF (I. Clément), ensuite saisis et munis d'un balisage minimal à Trèves (R. Töbner), à l'aide du logiciel TUSTEP. Les volumes suivants sont en attente d'un financement.

La deuxième étape s'effectue à l'Université de Liège, à l'aide du logiciel développé lors de l'étude de faisabilité (P. Renders et C. Briquet). Chaque article soumis au logiciel fait d'abord l'objet d'un prétraitement, destiné à détecter d'éventuelles erreurs de saisie ou coquilles typographiques. Le noyau du logiciel est constitué d'une trentaine d'algorithmes, qui détectent et balisent chacun un type particulier d'information lexicographique (étymon, renvoi, date etc.), en utilisant divers critères d'ordre typographique, structurel (positionnement par rapport à d'autres types d'information déjà identifiés) ou textuel (listes de mots-clés, expressions régulières). Enfin, des algorithmes dits de post-traitement facilitent la détection d'éventuelles erreurs de balisage et leur correction. L'ensemble du processus prend peu de temps : environ une seconde par article (138 secondes pour les 130 articles du corpus de test). Le balisage des volumes déjà saisis a commencé en août 2014 et doit s'achever dans le courant de l'année 2015 selon les volumes.

La mise en ligne des articles balisés, qui constitue la troisième étape du processus, s'effectuera en 2015 sur le site du FEW, hébergé à l'ATILF. L'interface d'exploitation est développée par l'équipe informatique de l'ATILF (E. Petitjean et B. Husson). Cette troisième étape nécessite en outre la création d'une police permettant d'afficher les caractères phonétiques du FEW (dont plusieurs ne sont pas encore pris en charge dans le standard Unicode) : l'élaboration de cette police de caractères a été confiée à l'Atelier National de Recherche Typographique (S. Kremer ; cf. http://www.anrt-nancy.fr/). Une proposition d'intégration de ces caractères dans le standard Unicode est en cours de rédaction.

En attendant l'arrivée des premiers volumes en mode texte, la totalité du FEW a été rendue disponible en mode image et pourvue d'une première interface de recherche, exploitant l'index sélectif du FEW (ATILF 2003). Au fur et à mesure de

l'avancement du chantier, les informations concernant le projet d'informatisation du FEW sont mises à jour sur les sites internet de l'ATILF (www.atilf.fr/few) et de l'Université de Liège (www.lingwa.philo.ulg.ac.be/recherche). Toute critique ou avis concernant le projet sont les bienvenus via les moyens de contact renseignés sur ces deux sites.

3. Fonctionnalités de recherche et de lecture

Afin de faciliter l'utilisation du FEW dans ses deux dimensions thesaurus et monographique, l'interface d'exploitation du FEW électronique proposera à la fois des fonctionnalités de recherche et des fonctionnalités d'aide à la lecture. Elle facilitera également la mise à jour de l'ouvrage et sa mise en réseau avec d'autres ressources lexicographiques et linguistiques en ligne.

Le balisage XML inséré dans le discours fewien permettra des recherches (simples ou combinées) sur divers types d'informations, parmi lesquels les étymons, les lexèmes et les informations qui y sont associées (signifiants, étiquettes géolinguistiques, catégories grammaticales, éléments de définition, sigles bibliographiques, datations), les mentions de langues romanes et non romanes, les mentions d'affixes ou encore les auteurs d'articles. Le FEW électronique exploitera en outre la base de données du *Complément*, afin de permettre des recherches plus précises sur des critères chronologiques. La recherche d'informations données de façon moins systématiques dans le discours fewien (telles que des caractéristiques morpholexicales par exemple) est envisagée dans un second temps.

Le résultat d'une recherche mènera nécessairement aux articles du dictionnaire : rappelons qu'une bonne utilisation du FEW, que ce dernier se présente dans une version imprimée ou dans une version électronique, « comporte une lecture (au moins cursive) de l'article complet dont relève la lexie à laquelle on s'intéresse » (ATILF 2003, VII). Il est en effet essentiel, pour s'approprier la totalité de l'analyse apportée par le FEW sur un

lexème, de replacer ce dernier dans son contexte. Diverses aides faciliteront la lecture d'un article :

- la mise en évidence des informations recherchées ;
- l'affichage d'un plan de l'article (généré automatiquement grâce au balisage XML) ;
- la mise en relation des marqueurs alphanumériques avec les références qui y renvoient dans le commentaire ;
- la mise en relation des notes et appels de note ;
- l'explicitation des sigles et abréviations (via l'accès électronique au Complément) ;
- la mise en place de liens hypertextes vers d'autres ressources lexicographiques en ligne.

Les modalités de mise à jour du FEW et de sa mise en réseau avec d'autres ressources en ligne sont en cours d'étude (cf. Renders 2014). Le FEW électronique doit faciliter l'accès aux nombreux compléments apportés au FEW depuis sa parution, que ces derniers concernent un article complet, une partie d'article ou une unité lexicale particulière. Les solutions proposées seront mises en pratique à la fois dans l'interface de lecture (accès aux compléments relevant de l'article consulté) et dans l'interface de recherche (possibilité de recherches à partir d'informations mises à jour). La question de la mise en réseau du FEW avec d'autres ressources informatisées (DEAF, TLFi etc.) est directement liée à celle de sa mise à jour ; les deux questions se résoudront notamment via la création d'url pérennes pour chaque article du FEW, voire, si cette proposition s'avère pertinente, pour chaque unité lexicale.

5. Bibliographie

ATILF, *Französisches Etymologisches Wörterbuch. Index A-Z*, Paris, Champion, 2003.

Büchi, Eva, *Les Structures du* Französisches Etymologisches Wörterbuch. *Recherches métalexicographiques et métalexicologiques*, Tübingen, Niemeyer, 1996.

CHAUVEAU, Jean-Paul, *D'un site informatique en chantier pour le FEW*, in : *Nuovi media e lessicografia storica. Atti del colloquio in occasione del settantesimo compleanno di Max Pfister*, Tübingen, Niemeyer, 2006, 33-37.

Complément = Chauveau, Jean-Paul/Greub, Yan/Seidl, Christian, *Französisches Etymologisches Wörterbuch. Eine darstellung des gallo-romanischen sprachschatzes. Complément*, Strasbourg, Éditions de linguistique et de philologie (Bibliothèque de Linguistique Romane, Hors Série 1), ³2010.

TLFi = CNRS/Université Nancy2/ATILF, Trésor de la langue française informatisé (cédérom), Paris, CNRS Éditions (version internet : http://stella.atilf.fr/), 2004.

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

RENDERS, Pascale, *Mise en ligne, mise à jour et mise en réseau du Französisches Etymologisches Wörterbuch*, in Trotter, David/Bozzi, Andrea/Fairon, Cédric (edd.), *Actes du XXVII^e Congrès International de linguistique et philologie romanes (Nancy, 15-20 juillet 2013). Section 16 : Projets en cours ; ressources et outils nouveaux*, Nancy, publication électronique, à paraître.

WOOLDRIDGE, Terence Russon, *Le FEW et les deux millions de mots d'Estienne-Nicot : deux visages du lexique français*, TraLiPhi 28 (1990), 239-316.

Pascale RENDERS