Trotter, David/Bozzi, Andrea/Fairon, Cédrick (éd.) (2016). *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16 : Projets en cours ; ressources et outils nouveaux*. Nancy, ATILF : http://www.atilf.fr/cilpr2013/actes/section-16.html.

Mise en ligne, mise à jour et mise en réseau du *Französisches Etymologisches Wörterbuch*

La mise en ligne et la mise en réseau des ressources lexicographiques constituent un chantier actuellement très important en Europe : le CNRTL[1] à Nancy, le projet GiGaNT[2] à Leiden ou le projet Wörterbuchnetz[3] à Trèves en sont quelques exemples. L'un des moyens mis en œuvre consiste à 'rétroconvertir' un dictionnaire imprimé en dictionnaire électronique en y insérant, de façon semi-automatisée, des balises XML. Cette opération s'avère particulièrement délicate dans le cas d'ouvrages aux structures complexes, dont on veut rendre compte sans en cacher les imperfections.

Parmi les ouvrages à informatiser de cette manière se trouve le FEW, œuvre incontournable dans le domaine de la linguistique française et romane et en même temps reconnue comme difficilement exploitable. Il est depuis longtemps acquis que l'informatisation du FEW, à condition de respecter les structures et la complexité de l'oeuvre, permettra de résoudre les problèmes d'accessibilité et ouvrira de nouveaux itinéraires de consultation. Par ailleurs, le FEW constitue un « dizionario-tetto » (Buchi/Renders 2013) qui serait bénéfiquement mis en réseau avec des dictionnaires en ligne tels que le *Dictionnaire étymologique de l'ancien français* (DEAF) ou l'*Anglo-Norman Dictionary* (AND).

La présente contribution a pour objectif premier de faire le point sur l'avancement du projet d'informatisation des 25 volumes du *Französisches Etymologisches Wörterbuch* (FEW). Nous présentons ensuite brièvement un projet récent, fortement lié au précédent et soutenu en Belgique par le Fonds National de la Recherche Scientifique sous l'intitulé « Étude des modalités de mise à jour et mise en réseau du *Französisches Etymologisches Wörterbuch*, dans le cadre de son informatisation, avec d'autres ressources linguistiques notamment belgoromanes ».

1. Informatisation du FEW et mise en ligne

L'informatisation du FEW et sa mise en ligne à travers une interface d'interrogation sont attendues depuis longtemps par la communauté scientifique : elles constituent ensemble la solution la plus efficace aux difficultés d'accès de l'ouvrage. L'objectif est double : d'une part, faciliter la recherche d'informations ciblées (étymons ou

[1] Cf. Pierrel / Buchi 2009 ; ‹www.cnrtl.fr/›.

[2] ‹www.inl.nl/onderzoek-a-onderwijs/projecten/gigant›

[3] ‹woerterbuchnetz.de›

lexèmes bien sûr, mais aussi références bibliographiques, suffixes etc.) ; d'autre part, faciliter la lecture d'un article grâce à diverses aides (plan de l'article, lien vers les abréviations des sigles, etc.).

Pour qu'un texte, quel qu'il soit, puisse être interrogé en ligne, trois étapes sont nécessaires : la mise en forme du texte, son enrichissement via des métadonnées et, enfin, la création d'outils de requête permettant d'interroger le texte via ces métadonnées. Il en va de même pour un dictionnaire : il s'agit (1) d'acquérir le texte (caractères spéciaux inclus) sous forme électronique ; (2) d'identifier sa structure et son contenu de façon explicite au moyen de métadonnées, de façon à permettre des requêtes sur les divers types d'information voulus ; (3) de donner accès au texte et aux métadonnées via une interface d'interrogation.

Idéalement, ces trois étapes se succèdent dans le temps, chacune permettant la suivante. En ce qui concerne le FEW cependant, la résolution de ces trois étapes n'a pas eu lieu dans l'ordre indiqué. Le problème de l'identification des informations a été étudié et résolu avant que ne soient trouvés des financements pour réaliser la première étape, qui est aussi la plus coûteuse. Le projet a donc été pensé de façon à ce que son avancement ne dépende pas de l'acquisition du texte entier des 25 volumes du FEW. Actuellement (juillet 2013), seule une acquisition partielle est en cours. La troisième étape reste, quant à elle, encore à réaliser. L'ensemble du projet fait l'objet d'une collaboration étroite entre l'ATILF (CNRS/Université de Lorraine) et le service de *Linguistique du français et dialectologie wallonne* de l'Université de Liège, dont nous faisons partie en tant que chargée de recherches F.R.S.-FNRS. L'acquisition du texte sous forme électronique bénéficie également de l'aide de la Fondation pour le FEW (Fondation pour le dictionnaire étymologique du français du professeur von Wartburg) et du Center for Digital Humanities de Trèves (http://kompetenzzentrum.uni-trier.de).

Les choix posés ne sont pas sans avoir d'influence sur la manière dont le FEW informatisé sera interrogé et utilisé. Nous expliquons ci-dessous l'avancement du travail et les solutions adoptées en ce qui concerne chacune de ces trois étapes, dans l'ordre qui est celui de leur résolution : l'identification des informations, l'acquisition du texte et, enfin, la création d'une interface.

1.1. Identification des informations

Si la deuxième des trois étapes précitées a été résolue en premier, c'est parce que l'ensemble du projet d'informatisation nécessitait une étude préalable de faisabilité, qui s'est clôturée en 2011 sous la forme d'une thèse de doctorat en cotutelle Liège-Nancy (Renders 2015). Les solutions adoptées pour l'identification des divers composants du discours 'fewien' sont le résultat de cette thèse, constituée d'une partie plutôt linguistique (métalexicographique) et d'une partie plutôt informatique (algorithmique). Il s'agissait en effet de proposer une modélisation du discours fewien avant d'étudier la façon dont cette modélisation, formalisée en XML, pourrait être intégrée de façon automatisée dans le texte lexicographique.

La modélisation du FEW, telle qu'elle a été proposée et adoptée, présente une caractéristique essentielle. Alors que la thèse d'Éva Buchi présentait les structures du FEW selon un point de vue interne (Büchi 1996), ces mêmes structures sont appréhendées, pour la modélisation, selon le point de vue de l'utilisateur. En effet, à partir du moment où l'objectif est de résoudre des problèmes d'utilisation, il nous paraît évident qu'il faut avoir une idée de l'identité des utilisateurs du dictionnaire, de ce qu'ils y cherchent, de la façon dont ils le consultent, des difficultés qu'ils rencontrent et des fonctionnalités qu'ils voudraient y trouver. Des discussions avec plusieurs spécialistes et avec les membres de la rédaction du FEW, ainsi que les réponses à un questionnaire distribué en 2007 à l'occasion du XXV^e^ CILPR, ont permis d'obtenir un aperçu des pratiques actuelles d'utilisation du FEW et de connaître les souhaits des utilisateurs dans l'optique d'un FEW informatisé, souhaits qui sont en relation étroite avec les difficultés qu'ils rencontrent dans la consultation de la version imprimée. L'étude a mis en évidence les parcours suivis par les utilisateurs au sein du dictionnaire et, surtout, deux modes d'utilisation de l'ouvrage, correspondant à deux dimensions du FEW : une utilisation du FEW comme thesaurus d'unités lexicales d'une part, comme recueil d'articles monographiques d'autre part. La vision du FEW comme thesaurus concerne notamment les parcours de consultation et les voies d'entrée dans le dictionnaire ; la vision du FEW comme recueil de monographies vise davantage les parcours de lecture et la mise en relation des informations au sein d'un article. Ces deux dimensions présentent une complémentarité essentielle, dans le sens où privilégier l'une des deux au détriment de l'autre est source d'approximations et d'erreurs. À partir de ce constat, nous avons pu construire une modélisation du FEW intégrant les deux dimensions précitées et respectant à la fois les structures du dictionnaire et les attentes de l'utilisateur. Le modèle identifie une trentaine de types d'information explicites ou implicites, situés à différents niveaux du FEW (de l'infrastructure à la superstructure, cf. Büchi 1996). La formalisation du modèle a été réalisée au moyen d'un schéma XML.

Il restait à trouver le moyen d'intégrer ce formalisme au sein du texte lexicographique de la façon la plus automatisée possible, un balisage manuel étant impossible du fait de la densité de l'ouvrage. C'était l'objectif de la partie algorithmique de la thèse, qui a effectivement conduit à l'automatisation complète du balisage. Cette automatisation est possible grâce à une séquence d'algorithmes qui, chacun, balise un des types d'information retenus (étymon, lexème, sigle bibliographique, datation, commentaire, champ documentaire, appel de note etc.). Les algorithmes de balisage utilisent des informations syntaxiques (fournies via les balises insérées par les algorithmes qui les précèdent dans la séquence) et des informations lexicales (fournies via des listes de mots-clés provenant, entre autres, du *Complément* au FEW). Un mécanisme de 'chaînes virtuelles' (cf. Briquet/Renders 2010) facilite la détection de ces informations en permettant de rendre invisibles certains éléments XML lors de la recherche de chaînes de caractères. Le tout s'intègre dans un logiciel dit 'de

rétroconversion' implémenté en java[4]. Les résultats fournis sont plus que satisfaisants, puisque la plupart des informations sont reconnues à cent pour cent dans le corpus de test, constitué d'une centaine d'articles concernés par le projet ANR DETCOL[5]. Les erreurs et oublis éventuels sont repérables grâce à un mécanisme d'alertes et de vérification visuelle intégré à la fin de chaque article XML produit. Enfin, le bon fonctionnement du logiciel ne nécessite pas que la totalité du FEW soit disponible : tout article ou échantillon d'articles fourni au logiciel, à condition qu'il soit muni d'un balisage minimal (uniquement typographique) et codé en Unicode, se voit en quelques secondes doublé d'une copie XML totalement balisée, dans laquelle toutes les informations voulues sont identifiées.

La modélisation du discours fewien et le logiciel de balisage permettent d'effectivement insérer dans chaque article du FEW les métadonnées adéquates : l'étude de faisabilité résoud ainsi la deuxième des trois étapes précitées. La mise en pratique nécessite toutefois que soit préalablement disponible une version électronique brute de l'article, ce qui nous ramène à la première étape, celle de l'acquisition du texte.

1.2. Acquisition du texte

L'opération consistant à acquérir une version électronique brute du texte lexicographique peut s'effectuer par saisie manuelle ou par océrisation, c'est à dire par l'application, sur un scan du document, d'un logiciel de reconnaissance optique. Quelle que soit la solution retenue, elle requiert également que soit choisie une police de caractères permettant le codage de tous les glyphes présents dans le dictionnaire.

Les tests d'océrisation réalisés sur le FEW ont fourni des résultats qui, malgré leur qualité (jusqu'à 96% de reconnaissance), nécessitaient ensuite un effort important de correction manuelle, plus coûteux en ressources temporelles et humaines que la solution d'une saisie directe. Cette dernière, proposée par le Center for Digital Humanities de Trèves, consiste à effectuer une double saisie, puis à comparer à l'aide d'un logiciel automatique les deux versions obtenues et à vérifier le résultat uniquement là où il y a divergence. Ce traitement offre un résultat excellent, voire parfait. Le coût pour la totalité du FEW s'élève à plusieurs dizaines de milliers d'euros. Heureusement, le logiciel de rétroconversion ayant été conçu de façon à pouvoir traiter le dictionnaire par ensemble d'articles, il est possible d'avancer en fonction des financements obtenus. Actuellement (juillet 2013), trois volumes, à savoir les volumes 16, 17 et 19, ont pu, avec l'aide de la Fondation pour le FEW, être scannés à l'ATILF et fournis au Center for Digital Humanities de Trèves, qui opère la double saisie et la vérification.

---

[4] L'implémentation en java du logiciel a été réalisée par Cyril Briquet, ingénieur de recherche à l'ATILF entre octobre 2008 et décembre 2009.

[5] Pour une présentation du projet ANR-DETCOL (Développement et Exploitation Textuelle d'un Corpus d'Œuvres Linguistiques), voir http://ctlf.ens-lyon.fr/documents/de_anr_workshop-corpus.asp

La question de la police de caractères est, quant à elle, directement liée à la présence de caractères spéciaux et à la question de leur codage : la police utilisée doit en effet contenir l'ensemble des caractères du texte. En ce qui concerne le FEW, un relevé a fait apparaître la présence de nombreux caractères phonétiques non standards, c'est-à-dire non répertoriés dans Unicode et, donc, non systématiquement pris en charge par les polices disponibles sur le marché. La présence de ces caractères non standards empêche d'une part l'acquisition, d'autre part la visualisation sur écran, du texte fewien. Pour lever le premier obstacle, il a été décidé de coder les caractères non standards en tant qu'entités XML (telles que « &few-e-ouvert-accent », par exemple). Cette solution permet de poursuivre l'informatisation du FEW depuis la phase d'acquisition jusqu'au balisage complet des articles. Seule la dernière étape du processus, à savoir la mise en ligne du dictionnaire et sa visualisation sur écran, nécessitent la création d'une police contenant l'ensemble des caractères fewiens. Le relevé de l'ensemble de ces caractères, accompagnés de leur codage Unicode (éventuellement dans la zone privée pour les caractères non standards) se trouve dans Renders 2015 ; ce relevé constitue une base technique pour la future police du FEW, dont la réalisation a été confiée en 2013 à l'Atelier National de Recherche Typographique (http://www.anrt-nancy.fr/).

1.3. Mise en ligne et interface d'interrogation

La dernière étape du processus, consistant à mettre en ligne les articles balisés et à les rendre interrogeables via un moteur de recherche, est prise en charge par l'ATILF. À l'heure où nous écrivons ces lignes, un prototype existe, mais il nécessite encore du développement. Soulignons que le FEW informatisé sera mis gratuitement à la disposition de la communauté des chercheurs : cette volonté, dont on peut se réjouir, s'inscrit dans le courant actuel de l'Open Access.

En attendant la création d'une interface d'interrogation en ligne, l'ATILF a prévu de rendre disponibles les scans en version image des volumes du FEW. Les trois volumes en préparation seront bientôt accessibles en format image via le site de l'ATILF (www.atilf.fr/few) ; les autres devraient suivre assez rapidement.

2. Mise à jour et mise en réseau

Parallèlement à l'acquisition, l'informatisation et la mise en ligne du texte complet du FEW, un nouveau projet de recherche a vu le jour, qui anticipe l'avenir et se penche sur deux questions : la mise à jour du FEW et sa mise en réseau avec d'autres ressources lexicographiques pour lesquelles il constitue une référence obligée. Ces deux problématiques sont liées, puisque la mise en réseau doit rendre disponible une partie des ajouts et corrections actuellement consignés dans des ouvrages annexes (cf. Greub 2012) ou dans des ressources en ligne (TLF-Étym, DEAF ou AND, pour ne citer qu'eux). Les possibilités d'informatisation de ressources belgoromanes corrigeant le FEW, telles que l'*Atlas Linguistique de la Wallonie* (ALW), font partie de l'étude.

Le corpus d'articles choisi pour tester les possibilités de mise à jour du FEW comprend notamment les articles du volume 19 (Orientalia), qui présentent l'intérêt d'avoir fait l'objet d'actualisations et de corrections systématiques (Arveiller 1999) et de pouvoir être mis en relation avec des notices du projet TLF-Étym consacrées aux arabismes du français (Baiwir 2010). Au-delà des possibilités et des choix informatiques, il s'agit d'étudier la façon dont la mise en relation du FEW avec des ressources en ligne susceptibles elles-même d'évoluer peut influencer les modalités de mise à jour du FEW lui-même.

Cette étude n'en est qu'à son début, mais elle a déjà permis de dégager des constats nets. La question de la mise à jour du FEW, tout comme celle de son informatisation et de sa mise en ligne, nécessite de cerner d'abord les parcours traditionnels suivis par les utilisateurs qui veulent vérifier l'actualité d'une information donnée par le FEW. Elle nécessite également que soit identifié le contenu possible de ces mises à jour : s'agit-il de compléments, de corrections, d'ajouts, de refontes ? Quelles parties du discours lexicographique sont concernées ? Ce n'est qu'ensuite que seront envisagées les possibilités de parcours informatiques, de façon à ce que le FEW informatisé permette les parcours traditionnels et les facilite. L'utilisateur pourrait par exemple, à partir d'un article en ligne, avoir accès à tous les compléments qui y sont apportés ailleurs, via des liens directs (uri, pour les ressources en ligne) ou via des références bibliographiques (pour les ressources disponibles uniquement en version imprimée). Dans le cas d'articles refondus, le parcours serait inversé : l'utilisateur partirait de l'article mis à jour et pourrait, s'il le désire, consulter l'ancienne version archivée. Rappelons que toute version d'un article, même obsolète, doit absolument être conservée, dans l'optique de permettre la vérification des sources d'une recherche : toute version est en effet susceptible d'avoir été utilisée et citée par un chercheur.

Nous pourrions peut-être aller plus loin encore et envisager une intégration des corrections et ajouts au sein même du FEW. En effet, la question de la mise à jour pourrait, en définitive, se résumer à ceci : les mises à jour du FEW peuvent-elles intégrer le FEW = avec le risque de perturber les structures du discours initial – ou doivent-elles rester à l'écart de l'ouvrage originel, si cohérent dans ses qualités et ses défauts? L'étude y répondra ; qu'il nous soit déjà permis d'affirmer qu'énoncée comme telle, et malgré sa réelle pertinence, cette question relève avant tout d'une conception 'papier' de l'objet dictionnairique. Par ailleurs, il ne sera jamais question de « mettre à jour le FEW », mais de mettre à jour, soit un article du FEW, soit une unité lexicale du FEW (pourvue de toutes les informations y associées) : ici comme précédemment, il est nécessaire, nous semble-t-il, de distinguer les deux dimensions du FEW, en tant que thesaurus et en tant que recueil de monographies. Cette distinction est peut-être la clé d'une mise à jour intégrée qui soit à la fois respectueuse des structures de l'ouvrage et susceptible de répondre malgré tout aux attentes des utilisateurs.

La question de la mise en réseau du FEW avec d'autres ressources informatisées devrait, quant à elle, se résoudre plus simplement, l'"adresse FEW' faisant partie du programme lexicographique de la plupart des ouvrages concernés. Il restera à définir pour le FEW informatisé des adresses pérennes qui puissent servir à la fois de référence et de lien hypertextuel.

3. Conclusion

Certes, le FEW n'est pas encore disponible en version électronique. Derrière ce constat désolant, se cache une grande activité : à l'heure où nous écrivons ces lignes, le FEW est scanné à Nancy, le texte de trois volumes est en cours d'acquisition à Trèves, le logiciel de balisage est opérationnel et attend d'être utilisé à Liège. Ces opérations devraient voir très bientôt leur achèvement. Si tout se passe comme prévu, l'année 2014 verra également la création de l'interface d'interrogation à l'ATILF et, espérons-nous, la création de la police de caractères par l'Atelier National de Recherche Typographique, ce qui permettra la mise en ligne effective des volumes 16, 17 et 19 du FEW. L'informatisation des volumes suivants dépendra de nouveaux financements, non encore obtenus. En attendant, nous espérons que la mise en ligne d'une partie du FEW permettra de fournir à la communauté scientifique un exemple, critiquable, de ce que peut être le FEW informatisé, mis à jour et mis en réseau.

FNRS / Université de Liège Pascale RENDERS

Références bibliographiques

ALW = Remacle, Louis / Legros, Élisée et al., 1953–. *Atlas Linguistique de la Wallonie. Tableau géographique des parlers de la Belgique romane d'après l'enquête de Jean Haust et des enquêtes complémentaires* (10 volumes), Liège.

AND = Rothwell, W. / Gregory, S. / Trotter, D. A. (dir.), 2005[2] [1977–1992[1]]. *Anglo-Norman Dictionary*. ‹www.anglo-norman.net›

Arveiller, Raymond, 1999. *Addenda au FEW XIX (Orientalia)*, Tübingen, Niemeyer.

Baiwir, Esther. *Refonte des notices 'étymologie et histoire' des articles suivants du Trésor de la Langue Française informatisé (TLFi) : achour (2010), alezan (2010), antari (2010), coufique (2010), fez (2010), ketmie (2010), amin (2010), sumac (2010)*. ‹www.atilf.fr/tlf-etym›.

Buchi, Éva / Renders, Pascale, 2013. « 46. Gallo-romance I : Historical and etymological lexicography », in : *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume : Recent developments with special focus on computational lexicography*, Grouws, Rufus et al. (ed.), De Gruyter, Berlin/New York.

DEAF = Baldinger, Kurt et al., 1971–. *Dictionnaire étymologique de l'ancien français*, Québec/Tübingen/Paris, 1971–. ‹http://www.deaf-page.de›

FEW = Wartburg, Walther von et al., 1922–2002. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes (25 vol.)*, Bonn / Heidelberg / Leipzig-Berlin / Bâle, Klopp / Winter / Teubner / Zbinden.

Briquet, Cyril / Renders, Pascale, 2010. « A virtualization-Based Retrieval and Update API for XML-Encoded Corpora », in : *Proceedings of Balisage: The Markup Conference 2010*, Balisage Series on Markup Technologies, vol. 5.

Buchi, Éva, 1996. *Les Structures du 'Französisches Etymologisches Wörterbuch'. Recherches métalexicographiques et métalexicologiques*, Tübingen, Niemeyer.

Complément = Chauveau, Jean-Paul / Greub, Yan / Seidl, Christian, 2010. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes. Complément*, Strasbourg, Éditions de linguistique et de philologie, Bibliothèque de Linguistique Romane, Hors Série 1.

Greub, Yan, 2012. « Sur le FEW », *The Aberystwyth Colloquium. Present and future research in Anglo-Norman (21-22 juillet 2011)*, Aberystwyth, Anglo-Norman Online Hub, 187-190.

Pierrel, Jean-Marie / Buchi, Éva, 2009. « Research and Resource Enhancement in French Lexicography : the ATILF Laboratory's Computerised Resources », in : Bruti, Silvia / Cella, Roberta / Foschi Albert, Marina (ed.), *Perspectives on Lexicography in Italy and Europe*, Newcastle upon Tyne, Cambridge Scholars Publishing, 79-117.

Renders, Pascale, 2015. *L'informatisation du Französisches Etymologisches Wörterbuch. Modélisation d'un discours étymologique*, Strasbourg, Éditions de linguistique et de philologie.

TLF-Étym = Collectif, 2005. *Projet TLF-Étym: mise à jour des notices étymologiques du Trésor de la langue française informatisé. Dossier de présentation*, Nancy, ATILF/CNRS/Université Nancy 2/UHP, ‹http://www.atilf.fr/tlf-etym› ; Voir aussi Buchi, Éva, 2005. « Le projet TLF-Étym (projet de révision sélective des notices étymologiques du Trésor de la langue française informatisé) », *Estudis romànics* 27, 569-571.