Pascale Renders

L'informatisation du FEW: quels objectifs, quelles possibilités?

0 Introduction

Le titre «informatisation du FEW» peut désigner de manière générale divers projets amorcés récemment au Centre du FEW (cf. www.atilf.fr/few). L'un d'eux, qui concerne la refonte des articles de la tranche alphabétique B-, est l'élaboration d'un outil d'aide à la rédaction (Matthey / Nissille, dans ce volume). Un autre projet, en cours d'étude à l'Université de Liège et à l'ATILF, consiste à informatiser les 25 volumes existants du FEW, dans le but de faciliter leur utilisation. C'est de cette *rétroconversion* du dictionnaire papier en dictionnaire électronique que nous parlerons ici. Nous ne nous éloignerons pas, ce faisant, des préoccupations de la section 13 du CILPR. D'abord parce que l'œuvre de von Wartburg représente une référence incontournable pour tout linguiste et philologue romaniste et que la question de sa consultation mérite donc d'être posée; ensuite parce que le FEW lui-même puise ses matériaux dans d'autres dictionnaires, glossaires et parfois éditions de textes qui sont en perpétuelle actualisation, ce qui pose la question de l'intégration de ces nouveaux apports dans les articles déjà rédigés, et donc celle de la mise à jour de l'ouvrage.

Une synthèse des données actuelles du problème permettra, tout d'abord, de définir les objectifs à atteindre et d'envisager une solution informatique (1). Nous nous interrogerons ensuite sur la faisabilité de l'opération, en abordant quelques-unes des problématiques que soulève chacune des étapes du processus (2).

1 Objectifs et méthode

1.1 Synthèse de la situation actuelle

L'utilisateur du FEW est actuellement dans une situation inconfortable. Il a devant lui un

ensemble de 25 volumes «bien tassés» dont la consultation n'est pas toujours commode. Repérer où se trouve une information demande un minimum de compétences en linguistique historique et un minimum de connaissances à propos de la structure du dictionnaire, lequel rassemble des milliers de renseignements de façon elliptique et souvent indigeste. Une fois l'information trouvée, il faut encore la comprendre, ce qui implique spécialement de résoudre correctement les abréviations et d'analyser exactement la place que l'information occupe dans la structure de l'article, notamment par la lecture du commentaire (généralement rédigé en allemand). Il est bien connu qu'en fin de compte, seule la pratique intensive de l'ouvrage procure à l'utilisateur une certaine «culture fewienne» qui peut rendre sa lecture moins fastidieuse.

Heureusement, le lecteur dispose de quelques outils précieux: les *Beiheft* et *Beiheft Supplement*, qui l'aident au décodage des abréviations, et l'index raisonné des formes du FEW (ATILF 2003), qui situe l'endroit du dictionnaire où est classé un lexème particulier. Ces outils, qui appartiennent à la péristructure du FEW, constituent des volumes supplémentaires à placer à côté du dictionnaire dans sa bibliothèque. De 25 volumes, nous passons à 29.

L'utilisateur consciencieux sait aussi que le FEW est sans cesse révisé et qu'il doit donc vérifier si l'information qu'il a trouvée et correctement interprétée n'a pas été remise en question depuis la publication, soit à l'intérieur même du dictionnaire (cf. Büchi 1996: 156-157), soit en dehors du FEW. Il explore, pour ce faire, la parastructure du FEW, constituée par une masse impressionnante de monographies, d'articles de revue et de comptes rendus qui proposent des ajouts et corrections à l'œuvre de von Wartburg. Il vérifie donc régulièrement la solidité des rayonnages de sa bibliothèque... si du moins il a la chance de posséder tous ces ouvrages et publications qui, sinon, sont consultables uniquement dans des bibliothèques spécialisées.

Le «cycle vertueux» que les responsables de la section 13 du CILPR ont évoqué concernant la lexicographie diachronique et la philologie est donc sérieusement mis à mal. D'une part, la masse de renseignements que contient le FEW est difficilement accessible et pose divers problèmes de lecture ou d'interprétation, même à des utilisateurs avertis. D'autre part, les nouveaux matériaux fournis par les linguistes et les philologues ne peuvent fusionner avec le FEW et restent confinés à une littérature annexe passablement dispersée.

1.2 Objectifs

1.2.1 L'informatisation du FEW a pour premier objectif de remédier à cette situation inconfortable. Il s'agit, d'une part, d'aider à la consultation du FEW, autant en ce qui concerne la lecture que la recherche des informations, d'autre part, de permettre l'actualisation de son contenu. Outre l'intérêt que présente la mise en ligne de ce dictionnaire, trois objectifs sont donc poursuivis:

1) *faciliter la lecture du texte*, notamment par l'aération de la mise en page, la résolution des abréviations, l'explicitation du plan des articles longs, la mise en évidence de certaines informations considérées comme particulièrement «utiles» et, de façon plus générale, par les fonctionnalités de navigation qu'offre l'outil

informatique;

2) *assurer l'exhaustivité d'une recherche*, non seulement en localisant l'endroit / les endroits du FEW où se trouve un lexème particulier, mais aussi en donnant accès aux étymons cachés et aux auto-corrections effectuées à l'intérieur même des colonnes du dictionnaire (cf. 1.1);
3) *permettre la mise à jour des données* et intégrer notamment au dictionnaire (selon des modalités à définir) les nombreuses corrections déjà publiées ou en cours.

1.2.2 Ce faisant, l'informatisation veut également rendre possible une *nouvelle utilisation du FEW*. Elle ouvre en effet la voie à des consultations «transversales» (effectuées à travers l'ensemble du dictionnaire) qui sont actuellement impossibles ou très ardues, du type «quels sont les mots du wallon liégeois hérités d'un étymon germanique et attestés avant 1800?». Elle permet aussi d'envisager la réalisation de «rêves fous», tels que la traduction en français des commentaires allemands, la réalisation d'une cartographie automatique (permettant de visualiser sur la carte du domaine galloroman les données localisées d'un article) ou encore la mise en relation du FEW avec d'autres documents, bases de données et dictionnaires informatisés.

1.3 Solution informatique

La solution informatique retenue s'appuie sur la méthode utilisée lors de l'élaboration du TLFi (cf. Dendien / Pierrel 2003): elle consiste à utiliser un format de balisage (dans notre cas XML) pour modéliser le discours du dictionnaire. Quatre étapes sont nécessaires, les trois premières conduisant au document informatisé: 1) l'acquisition du texte; 2) la définition d'un balisage rendant compte du discours étymologique propre au FEW; 3) la construction d'automates (programmes informatiques) permettant d'insérer ces balises aux bons endroits du texte; 4) l'exploitation du résultat via un moteur de recherche adéquat.

2 Faisabilité

Nous proposons de passer en revue ces quatres étapes et de présenter brièvement, pour chacune de celles-ci, les principales questions qui se posent, ainsi que les solutions actuellement envisagées.

2.1 Acquisition du texte

La première étape du processus d'informatisation, la plus déterminante car conditionnant les suivantes et donc le succès de toute l'entreprise, mais aussi la plus coûteuse (cf. Martin 2004: 145), est l'acquisition du texte brut du FEW dans un format électronique. Nous sommes ici, à peu de choses près, dans la situation de l'éditeur qui doit reproduire un texte

manuscrit: il s'agit tout d'abord de savoir quelles informations doivent être récupérées. Faut-il, par exemple, conserver une trace de la mise en page du texte du FEW en colonnes? Faut-il reproduire tous les caractères phonétiques dans leurs moindres détails? Les réponses dépendent de l'exploitation qui sera faite du texte par la suite: il est par exemple essentiel de conserver les données typographiques qui doivent servir d'indicateurs aux programmes de balisage (cf. 2.3).

Ces questions une fois résolues, il reste à savoir comment procéder, ce qui revient à choisir entre une numérisation et une saisie manuelle. La première est moins coûteuse, mais ne peut donner un résultat infaillible; il faudra ensuite rechercher toutes les erreurs pour les corriger manuellement. Des tests réalisés sur le FEW par Hassen Hadj Ammar (ATILF) donnent un résultat d'environ 96% de reconnaissance, ce qui laisse encore une marge d'erreurs non négligeable.

Une série d'articles ont déjà été numérisés par reconnaissance optique des caractères, puis corrigés manuellement pour servir de support aux tests de balisage. Signalons en outre que les caractères phonétiques spécifiques au FEW ont fait l'objet d'un relevé et que la création d'une police adéquate est en cours d'élaboration à Liège.

2.2 Modélisation du discours fewien

Par *modélisation*, nous entendons l'élaboration d'un balisage qui rende compte à la fois des divers types d'informations contenus dans le FEW et de leur structuration. Il s'agit de l'étape proprement «linguistique» du processus d'informatisation.

Le FEW présente un discours hautement construit. Cela signifie qu'il n'est pas possible de le réduire tel quel à une base de données: il faut garder trace de la façon dont les matériaux ont été agencés, car cette présentation est une analyse à elle seule. Le choix du format XML répond parfaitement à cette exigence, puisque ce dernier a pour principe de rendre compte de la structure du document indépendamment de sa présentation typographique. Depuis la parution de la thèse d'Eva Buchi sur les structures du FEW (Büchi 1996), il est envisageable de définir un modèle de balisage adéquat. En pratique, une série de paramètres compliquent néanmoins la tâche. Ils découlent pour la plupart d'une particularité majeure du discours fewien, à savoir le hiatus entre la structure de surface et la structure profonde de l'ouvrage.

Nous proposons ci-dessous deux exemples illustrant cet hiatus, pris l'un dans la microstructure, l'autre dans l'infrastructure du FEW.

2.2.1 Le balisage de la microstructure

Un principe fondamental du FEW est sa souplesse microstructurelle: chaque article présente les matériaux différemment selon les particularités de la famille lexicale traitée. Alors que dans l'informatisation du processus de rédaction, l'écueil à éviter est celui de contraindre le rédacteur à adopter une organisation hiérarchique préétablie, le problème est ici inversé: nous traitons des articles qui présentent une structure figée, et nous voulons

rendre compte de toutes ces structures différentes. Ceci serait tout à fait concevable si les critères de regroupement des formes étaient explicités en structure de surface et donc facilement reconnaissables. Malheureusement, si certaines étiquettes apparaissent avec plus ou moins de régularité – notamment pour les regroupements morphologiques sous des rubriques telles que *dérivés* et *composés* – la plupart des critères sont implicites. Ainsi, seule une lecture attentive des formes d'un paragraphe peut indiquer que le point commun les rassemblant est l'appartenance à une même famille dérivationnelle. Ce véritable casse-tête est dû à la combinaison de deux principes de base de la rédaction du FEW: la souplesse et l'économie.

Pour le moment et en attendant mieux, la solution retenue dans la modélisation de la microstructure consiste donc à rendre compte des regroupements des unités jusqu'à un certain niveau hiérarchique par un unique élément XML, en ne définissant le critère en présence que s'il est explicité en structure de surface.

2.2.2 Le balisage de l'infrastructure

Le même conflit entre structure de surface et structure profonde se pose pour l'infrastructure lorsqu'on envisage l'actualisation du FEW. La plupart des ajouts ou corrections à l'œuvre de von Wartburg concernent un lexème particulier: sa datation, son étymologisation («adresse FEW») ou même son existence à proprement parler (cf. «mots-fantômes»: http://www.atilf.fr/MotsFantomes). L'intégration dans le dictionnaire de ces corrections ponctuelles nécessite donc de reconnaître la «fiche» des rédacteurs (cf. Matthey / Nissille, dans ce volume), c'est-à-dire l'*unité minimale de traitement* (cf. Büchi 1996: 116). Malheureusement, celle-ci est détruite en structure de surface, puisque le FEW ne répète pas les informations qui ont déjà été citées pour une unité précédente. Cela complique considérablement la mise à jour du FEW par simple ajout, suppression ou déplacement d'une information, car de telles opérations sont susceptibles de nuire à l'intégrité des unités adjacentes.

Heureusement, il est en principe possible – l'inverse est dû à des «fautes» de rédaction – de reconstruire de façon automatisée les unités minimales de traitement en parcourant la suite des données présentées en structure de surface et en rétablissant les informations élidées. Cette opération permet de baliser chacune des unités présentes dans une séquence de texte et, donc, de faire réapparaître la structure profonde du dictionnaire.

Par rapport à la rédaction du FEW, le processus se trouve donc inversé, mais le principe est le même, à savoir la primauté donnée à la «fiche». Il est normal que l'on rejoigne ici les solutions que les rédacteurs ont adoptées dans l'informatisation de leur travail (cf. Matthey / Nissille, dans ce volume), puisque la mise à jour des matériaux oblige à imaginer un mécanisme de réécriture du FEW. L'informatisation du processus de rédaction nécessite une modélisation XML souple avant de pouvoir figer l'article dans sa version finale; de même, l'actualisation des matériaux exige, tout en conservant une trace de l'agencement des données tel qu'il a été décidé par le rédacteur, de pouvoir déconstruire l'article (le ramener à un ensemble de fiches) pour effectuer les modifications voulues avant de rétablir l'implicite.

2.3 Écriture des algorithmes de balisage

Une fois défini, le balisage peut être intégré dans le texte. Le contenu volumineux du FEW rendant inconcevable une opération manuelle, la troisième étape de l'informatisation consiste à écrire des automates qui insèreront les balises aux bons endroits. La conception des automates nécessite la définition préalable d'*indicateurs de reconnaissance* (des indices permettant d'identifier de façon fiable une information), qui peuvent être de différents types: typographiques (par ex. le gras indiquant un étymon vedette), textuels (par ex. l'appartenance d'une séquence à la liste des sigles bibliographiques répertoriés dans le *Beiheft*) ou structurels (par ex. la situation de la glose à la fin du champ de l'entrée).

Le succès des opérations dépend majoritairement de deux paramètres: la présence d'indicateurs en nombre suffisant et leur fiabilité d'un bout à l'autre du document à traiter. Dans le cas du FEW, la tâche apparaît plus ardue encore que pour le TLF. En effet, d'une part, le nombre d'indicateurs est mis en péril par le principe d'économie cher au FEW, d'abord à cause de l'implicite, qui rend malaisée la reconnaissance de certaines informations, ensuite parce que des indices ténus sont dotés d'une grande importance: la ponctuation, par exemple, joue un rôle essentiel dans le regroupement des formes, et le moindre séparateur non reconnu est un facteur d'erreur dans l'interprétation fine du texte. D'autre part, la fiabilité des indicateurs pose elle aussi problème, le FEW présentant une assez grande instabilité d'un volume à l'autre et d'un article à l'autre. Les nombreuses irrégularités (cf. Büchi 1996: 129) et les erreurs (absence de parenthèse fermante, coquilles etc.) sont à même de contrarier l'informatisation de l'ouvrage.

Heureusement, certaines informations s'avèrent directement et facilement repérables: citons notamment l'étymon vedette, les étiquettes géolinguistiques, les sigles bibliographiques, les formes ou encore les définitions. Leur balisage sert donc de point de départ pour l'identification des autres objets textuels. L'échec de reconnaissance de certains d'entre eux, en raison de l'absence d'indicateurs ou d'une trop grande irrégularité, peut enfin obliger à apporter des aménagements au modèle élaboré à l'étape précédente (2.2). Un dialogue est donc nécessaire entre la modélisation souhaitée et les possibilités pratiques de reconnaissance des informations à baliser.

2.4 Exploitation du résultat: moteur de recherche et interface de requête

La dernière étape du processus ne retiendra pas notre attention ici, car elle ne dépend pas du FEW lui-même et ne pose pas de problème majeur. Une fois le FEW informatisé, il «suffira» de l'exploiter en proposant au lecteur, probablement sur Internet, une interface de consultation et de recherche plus ou moins semblable à celle qui existe pour le TLFi. C'est ici qu'interviendront les relations entre le FEW et d'autres bases et dictionnaires informatisés, ainsi qu'avec les outils annexes actuellement en développement à l'ATILF, tels que le *Beiheft*, converti en base de données, ou l'index onomasiologique.

3 Conclusion

L'informatisation du FEW apporterait une solution pleinement efficace aux difficultés de consultation de l'ouvrage, difficultés sur lesquelles de nombreux linguistes et philologues ont déjà attiré l'attention. Elle représente en outre le seul moyen d'envisager une mise à jour pratique et accessible. Cette informatisation est attendue depuis longtemps; déjà en 1990, T. R. Wooldridge (1990: 239) affirmait:

> La rédaction du *FEW* s'achève; la tâche immense mais logique qui s'impose dorénavant est de l'informatiser, en y intégrant, sans briser la méthode mais en permettant d'autres lectures des données, toutes les corrections et compléments sûrs qui ont été apportés depuis plusieurs décennies et en y ajoutant ceux à venir.

Si l'intérêt du projet est évident, la question de sa faisabilité reste toutefois délicate. Un constat s'impose: rien ne sera possible sans la réussite de la première étape du processus, à savoir l'acquisition du texte fewien dans un format électronique, par numérisation ou par saisie manuelle. Cette étape essentielle une fois franchie, nous savons d'ores et déjà qu'un balisage minimal sera réalisable sans trop de difficultés. La lecture de l'ouvrage sera donc bel et bien facilitée, de même que la recherche simple d'éléments apparaissant explicitement en structure de surface. Cela rendrait déjà d'immenses services à la communauté en résolvant une grande partie des difficultés actuelles.

Pour effectuer des recherches plus complexes et exploiter au mieux l'œuvre de von Wartburg et de ses héritiers, un étiquetage plus fin du discours fewien est nécessaire. Celui-ci se heurte malheureusement à divers problèmes qui relèvent de l'inaccessibilité de la structure profonde du FEW. Or, c'est cette dernière qui intéresse les utilisateurs.

Partant du principe qu'un projet informatique n'est valable que s'il rencontre les besoins du «client» –dans notre cas, les utilisateurs actuels et futurs du FEW–, nous énonçons donc un deuxième principe, à savoir la nécessité de modéliser le dictionnaire en cherchant à atteindre sa structure profonde. Ces deux principes sont résumés dans un troisième et dernier, qui sous-tend l'ensemble de nos réflexions: l'informatique doit se plier aux exigences de la linguistique – et non l'inverse. Nous tâcherons donc de rendre compte du discours fewien dans ses particularités, sa diversité et sa profondeur. Nous tenterons également de trouver des solutions pour répondre aux attentes légitimes exprimées par les utilisateurs du FEW, ne nous résignant à baisser les bras que devant le bon sens, qui ne peut quitter aucune entreprise, si ambitieuse soit-elle.

Bibliographie

ATILF (2003): *Französisches Etymologisches Wörterbuch. Index A-Z* (2 voll.). Paris: Champion.

Baldinger, Kurt (1974): *Le FEW de Walther von Wartburg. Introduction*. In: Baldinger, Kurt (ed.): *Introduction aux dictionnaires les plus importants pour l'histoire du français*. *BJR* 18-19,

Strasbourg (= Paris: Klincksieck), 11-47.
— (1988-2003): *Etymologien. Untersuchungen zu FEW 21-23* (3 voll.). Tübingen: Niemeyer.
Beiheft = Wartburg, Walther von (^{2}1950): *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes. Beiheft: Ortsnamenregister, Literaturverzeichnis, Übersichtskarte*. Tübingen: Mohr.
Beiheft Supplement = Hoffert, Margarete (1989 [1957]): *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes. Supplement zur 2. Auflage des Bibliographischen Beiheftes*. Bâle: Zbinden.
Büchi, Eva (1996): *Les Structures du* Französisches Etymologisches Wörterbuch. *Recherches métalexicographiques et métalexicologiques*. Tübingen: Niemeyer.
Dendien, Jacques / Pierrel, Jean-Marie (2003): *Le Trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence*. In: *Traitement automatique des langues* 43, 2, 11-37.
Martin, Robert (^{2}2004): *Comprendre la linguistique. Épistémologie élémentaire d'une discipline*. Paris: Quadrige / *PUF*.
Renders, Pascale (en préparation): *Modélisation d'un discours étymologique. Prolégomènes à l'informatisation du FEW*. Liège: Université de Liège [thèse de doctorat].
Wartburg, Walther von (1961): *L'expérience du FEW*. In: *Lexicologie et lexicographie françaises et romanes. Orientations et exigences actuelles* (Strasbourg 12-19 novembre 1957). Paris: Éditions du CNRS, 209-219.
Wooldridge, Terence Russon (1990): *Le FEW et les deux millions de mots d'Estienne-Nicot: deux visages du lexique français*. In: *TraLiPhi* 28, 239-316.