MethIS, 2 (2009), p. 179–195.

Des dangers de l'informatisation d'un document : le cas du *FEW*

P. Renders

(Université de Liège)

Résumé. La plupart des travaux de recherche en linguistique historique du français nécessitent l'exploitation d'un outil de travail spécifique à la discipline, le *Französisches Etymologisches Wörterbuch*. Les structures complexes de ce dictionnaire le rendent atypique dans l'ensemble des productions lexicographiques et entraînent des pratiques particulières qui restreignent ses possibilités d'utilisation. Son informatisation devrait faciliter sa consultation et permettre de nouvelles modalités d'interrogation. Elle induit toutefois des « dangers » pour la réception du document, qui se situent à la fois dans la manière dont ce dernier est traité lors du processus et dans la façon dont le résultat sera exploité. En fin de compte, l'informatisation peut être assimilée au processus de traduction : tous deux sont à appréhender avec autant d'enthousiasme que de prudence.

1. Introduction

À l'ère de l'informatique, le fait d'émettre des réserves sur l'intérêt de numériser un document ressemble sans doute à de la provocation. En posant, entre autres, la question de ce nouveau support comme défi aux pratiques du document, le présent volume soulève pourtant un problème qui ne peut être écarté trop rapidement. L'enthousiasme que suscitent les innovations apportées par l'informatique dissimule les conséquences dommageables qu'elles peuvent entraîner sur la réception des documents et sur leur bonne exploitation.

Les réflexions qui suivent s'ancrent dans une étude préalable à l'informatisation du *Französisches Etymologisches Wörterbuch*[1]. Nous n'avons nullement la prétention de généraliser notre propos, d'autant plus qu'un dictionnaire est

1 Cette étude est réalisée dans le cadre d'une thèse de doctorat à l'ULg, en collaboration avec l'ATILF (CNRS/Nancy-Université).

un document tout à fait particulier. Le *FEW* a toutefois l'avantage de poser un certain nombre de questions intéressantes. En lexicographie, le bouleversement dû à l'irruption de l'informatique est réel et a déjà fait l'objet de nombreuses études[2]. Certains dictionnaires numérisés n'ont de neuf que le médium électronique et restent conçus dans l'optique d'une consultation tout à fait « traditionnelle » :

> *Encarta*, en ce sens, reste un dictionnaire assez proche, dans sa conception, du dictionnaire sur papier […]. Son utilisation offre essentiellement l'intérêt d'être plus maniable et plus rapide que celle qu'offrirait une version sur papier. (Ch. Jacquet-Pfau 2005 : 57)[3]

Cependant, un certain nombre de fonctionnalités propres à l'outil électronique peuvent générer une nouvelle pratique du document :

> *The true core of the ED [electronic dictionary] revolution however, lies in the fact that users are liberated from the alphabetical straitjacket, that hypertext, menus, etc. eliminate (artificial) linear text restrictions, that the data conjured up onscreen are not static, and that powerful search capabilities ensure a smooth overarching navigation*. (G.-M. de Schryver 2003 : 157)

La navigation hypertextuelle, les liens avec d'autres dictionnaires ou bases de données ou encore la présence d'outils divers parfaitement intégrés au document électronique permettent d'autres modes de lecture ; on notera en particulier la possibilité de recherches transversales (portant sur l'ensemble de l'ouvrage) et multicritères (R. Martin 1999 : 58). Grâce à son informatisation, l'ouvrage peut en outre devenir évolutif (intégration des mises à jour) et multimédia[4].

Ce changement de pratique est généralement, et avec raison, vu comme positif. Pourtant, à la suite d'une enquête réalisée en 1996 auprès des spécialistes du *FEW*, le directeur du *Dictionnaire Étymologique de l'Ancien Français* Frankwalt Möhren a mis en garde contre certaines des fonctionnalités que

2 Voir par exemple, dans le domaine francophone, B. Quemada (1991) et R. Martin (1999) ou, dans le domaine anglophone, G.-M. de Schryver (2003).

3 « […] *without actually doing something about the contents too (through the addition of more and new types of information), and without truly implementing fully integrated hypermedia access structures, EDs aren't really very different from their paper counterparts* […] » (G.-M. de Schryver 2003 : 146).

4 Pour un relevé plus complet des nombreux avantages qu'offrent les dictionnaires électroniques, voir G.-M. de Schryver (2003 : 155-160).

pourrait offrir l'informatisation, en signalant que les utilisateurs allaient « faire des bêtises ». Lesquelles, pourquoi ? Après une brève présentation de l'œuvre (2) et des pratiques qui accompagnent son état imprimé (3), nous étudierons les risques que comporte l'informatisation du *FEW* à deux niveaux : dans le traitement du document et dans l'exploitation du résultat (4).

2. Le *FEW* : un document particulier

Le premier fascicule du *Französisches Etymologisches Wörterbuch* a été publié en 1922 par Walther von Wartburg, non seulement dans l'optique que suggérait son titre, mais aussi dans le but plus ambitieux d'établir un thesaurus exhaustif du lexique galloroman, comme l'indique le sous-titre de l'ouvrage (*Eine Darstellung des galloromanischen Sprachschatzes*). Cette œuvre monumentale compte aujourd'hui 25 volumes[5], qui rassemblent toutes les lexies du français, du francoprovençal, de l'occitan et du gascon ainsi que de tous leurs dialectes, depuis les premières attestations (842) jusqu'à l'époque contemporaine.

Ces millions d'unités ont été analysées, regroupées et hiérarchisées de manière raisonnée selon leur histoire, justifiant ainsi l'adjectif « étymologique » de l'intitulé de l'œuvre. Le lemme d'un article du *FEW* est en effet l'étymon latin, germanique ou autre duquel proviennent, par transmission héréditaire ou par emprunt, direct ou indirect, toutes les lexies rassemblées dans l'article[6]. Ces dernières sont classées de manière à retracer précisément l'histoire de la famille lexicale ainsi reconstituée, selon une structure assez complexe et, de plus, variable en fonction des spécificités de chaque cas. Un commentaire explicatif, servant notamment à justifier le classement, suit la présentation des matériaux. Chaque article constitue donc une monographie où sont présentés, discutés et établis les liens historiques qu'entretiennent toutes les unités d'une même famille lexicale. Dans l'article TERROR par exemple (*FEW* 13/1, 264b-265a), le classement dans le paragraphe I 3 du dérivé *terroriste* explique à la fois les conditions de son apparition et son sémantisme, étroitement liés à la Révolution française :

5 La refonte du premier volume, dont les volumes 24 et 25 constituent la première partie (tranche alphabétique *A-*), se poursuit actuellement sous forme d'articles électroniques publiés sur Internet (cf. http://www.atilf.fr/few).

6 Dans la partie regroupant les matériaux d'origine inconnue, c'est toutefois le concept, et non l'étymon, qui fait figure de mot-vedette (cf. É. Büchi 1996 : 40 ; 79).

terror épouvante.

I. 1. Apr. *terror* f. „émotion violente, faite d'effroi et d'appréhension" (13e–15e s.), mfr. frm. *terreur* (dp. env. 1356), dauph. *terrour*, pr. *tarrour*,

(...)

2. Apr. *terror* „menace grave, intimidation (t. jur.)" (Narbonne 1254), mfr. *terreur* (mil. 15e s.).

3. Frm. *terreur* „ensemble de moyens de coercition politique, maintenant les opposants dans un état de crainte et brisant toute résistance" (dp. 1789, Br 9)[2]), spécialt. *la Terreur* „le régime instauré en France entre septembre 1793 et juillet 1794 (t. d'hist.)" (dp. 1794, Br 9)[3]). — Dér. Frm. *terroriste* m. „partisan, agent d'un régime de terreur" (dp. 1794, Br 9; ‚inus.' Zm)[4]), *terrorien* (1797, Br 9), *terroriste* „celui qui lutte clandestinement, par des moyens violents et criminels, contre un régime politique" Zm[5]), *terrorisme* „régime de terreur politique" (dp. 1794, Br 9; ‚inus.' Zm)[6]),

(...)

Emprunté de lt. TERROR par la langue lettrée (I 1), fr. *terreur*, utilisé parfois dans le style jur. (I 2), reçut de la Révolution une acception qui le fit entrer dans le vocabulaire politique (I 3). — Lt. *terrificus* a été emprunté passagèrement au XVIe s. (II 1 a) et une fois par les Goncourt (b). Le verbe lt. *terrificare* „répandre la terreur" a été empr. une première fois au 16e s. dans la forme du part. prés. (2 a), puis à la fin du 18e s. comme verbe. Voir encore TERRERE. — Zumthor.

La répertorisation exhaustive du lexique d'un ensemble de langues, la présentation raisonnée des matériaux et la mise en pratique aboutie du concept d'étymologie-histoire du mot (opposée à l'étymologie-origine pratiquée jusqu'alors) font du *FEW* un document unique et irremplaçable, qui reste une référence non contournable pour tout travail de recherche en linguistique historique française et romane.

3. Pratiques du *FEW* imprimé

Les caractéristiques du *FEW* le rendent atypique parmi l'ensemble des productions lexicographiques. Dans un dictionnaire étymologique classique, la nomenclature correspond aux données connues (les lexies du français moderne dont on cherche l'étymon), les informations historiques apparaissant dans le corps de l'article. À l'inverse, le *FEW* présente en mot-vedette l'étymon

lui-même. Ce classement se révèle évidemment peu utile lorsque c'est justement l'étymon que l'on recherche :

> Pour chercher l'origine d'une forme lexicale dans le *F.E.W.*, il faut connaître la nature « romane », « germanique » ou « empruntée » de son étymon et avoir des notions de phonétique historique [...]. L'étymon trouvé, il faut repérer dans un ou plusieurs sous-ensembles la forme recherchée, ce qui n'est pas toujours commode dans les grands articles [...]. (A. Rey 1971 : 103-104)

Plus généralement, un grand nombre de données contenues dans le dictionnaire s'avèrent difficilement accessibles, qu'il s'agisse d'étymons constitués en « sous-lemmes » peu visibles, de formes classées à plusieurs endroits de l'ouvrage (« étymologies doubles »[7]) ou de corrections internes effectuées dans les commentaires ou les notes des articles. Or, le *FEW* ne se consulte pas seulement pour étymologiser, dater ou localiser un lexème particulier, mais aussi comme corpus pour l'étude de phénomènes morphologiques, sémantiques ou phonétiques. Si l'on veut assurer l'exhaustivité et la pertinence de telles recherches, la seule « bonne pratique » reste, si l'on en croit un rédacteur formé par Wartburg lui-même, de « lire le *FEW* d'un bout à l'autre » (K. Baldinger 1974 : 25). Sagement, les utilisateurs du dictionnaire se contentent de lire les seuls articles qu'ils ont sélectionnés, ce qui s'assimile déjà en soi au parcours du combattant. À la compréhension des critères de classement des formes — que le commentaire n'expose pas nécessairement dans le détail — s'ajoute en effet le déchiffrage de nombreuses abréviations, résolues dans un recueil annexe[8]. La nécessité d'être accompagné de divers répertoires, dont un index sélectif des lexèmes traités[9], est paradoxale pour un dictionnaire et, de ce fait, révélatrice : dans les pratiques qui accompagnent son état imprimé, le *FEW*, malgré son apparence et son intitulé, est moins à appréhender comme un outil lexicographique que comme un ensemble de « textes » que le lecteur doit déchiffrer, comprendre, analyser et interpréter. Il en découle que seuls les initiés consultent le *FEW*, qui reste hermétique pour qui n'a pas été formé à l'exercice.

---

7 Voir par exemple K. Baldinger (1980).

8 Il s'agit du *Beiheft* (W. von Wartburg 1950²) et du *Beiheft Supplement* (M. Hoffert 1989), actuellement en cours de révision sous la direction de Jean-Paul Chauveau.

9 ATILF 2003. Avant la parution de cet index raisonné des formes du *FEW*, d'autres dictionnaires faisaient fonction d'entrée dans le corps de l'ouvrage.

4. « Dangers » de l'informatisation

De ces difficultés de consultation résulte une sous-exploitation du dictionnaire. Afin d'y remédier, les linguistes se sont récemment tournés vers l'informatique. La « rétroconversion » du *FEW* en dictionnaire électronique permettrait non seulement de résoudre les problèmes dus à sa complexité, mais aussi de le mettre à jour et d'offrir de nouveaux modes de consultation. Ces innovations permises par l'informatique — et souhaitables — comportent toutefois certains risques de mauvaise exploitation. D'autre part, et avant même d'aborder cette phase ultime du processus, le traitement effectué sur le document comporte une série d'opérations délicates qui méritent réflexion.

4.1. Traitement du document

Informatiser un document consiste d'abord — et parfois uniquement — à le rendre accessible sous forme électronique, par saisie manuelle ou par océrisation[10]. Cette opération pose déjà quelques problèmes pour le *FEW*, notamment quant à la conservation des notations phonétiques détaillées qu'il contient sous forme de caractères spéciaux. Il peut s'agir ensuite de rendre le document plus ou moins finement « interrogeable », par exemple en y insérant un balisage XML. Les informations sont alors étiquetées en fonction de leur type (une définition sera par exemple précédée d'une balise ouvrante <def> et suivie d'une balise fermante </def>), et leur structuration est explicitée. Ces annotations permettront à l'utilisateur d'interroger le texte de façon ciblée et d'extraire de l'ensemble de l'ouvrage les séquences pertinentes en fonction de ses critères de recherche.

Différents étiquetages d'un même document sont évidemment possibles, en fonction du point de vue que l'on adopte sur celui-ci et des informations qu'on désire y trouver. Cette affirmation est valable même dans le cas d'un dictionnaire, dont la structure rigoureusement codifiée prête, *a priori*, peu à discussion. L'automatisation du traitement influence également le modèle élaboré, en interdisant, au moins provisoirement, les balises qui ne peuvent être insérées sans intervention manuelle. Or, les choix réalisés à ce moment du processus ne seront pas sans conséquences pour la bonne exploitation du document et de son contenu : « […] les outils de recherche, associés à la

10 Voir P. Renders (à paraître) pour plus de détails sur la méthodologie suivie dans la rétroconversion du *FEW*.

structuration du dictionnaire, en font toute la richesse… ou toute la pauvreté ! » (Ch. Jacquet-Pfau 2005 : 67). Une première tentative de modélisation du *FEW* sous la forme d'un balisage XML a été réalisée en 2004 à l'ATILF dans le cadre de la refonte des articles de la tranche alphabétique B-. Malgré sa qualité, ce balisage s'est rapidement révélé inadéquat à divers endroits. Les problèmes posés peuvent être mis en relation avec deux démarches préalables à l'établissement d'un balisage : d'une part, l'analyse structurelle du document ; d'autre part, la normalisation des irrégularités qu'il présente.

4.1.1. *Analyse structurelle*

Une première opération nécessaire avant de baliser un document consiste à *établir le relevé des divers types d'informations* qu'il contient. Dans un dictionnaire, si certaines sont évidentes (le mot-vedette, la définition, la catégorie grammaticale, etc.), d'autres le sont moins et risquent d'être oubliées malgré leur intérêt. Les informations morphologiques telles que les suffixes et les préfixes ont par exemple été laissées de côté dans la première version du balisage, parce qu'elles n'apparaissaient pas de façon systématique dans le *FEW*. Or, elles intéressent beaucoup les utilisateurs du dictionnaire. La structuration des données a également été mal réalisée à certains niveaux, ce qui est plus grave : en définissant, par exemple, dans chaque paragraphe deux éléments distincts « Lexicographie française » et « Dialectes », le balisage insinue que le texte fewien fait une distinction entre les informations non dialectales et dialectales, ce qui est fondamentalement contraire à la vision de l'étymologie telle qu'elle se dégage du *FEW*. Un tel traitement est dangereux parce qu'il propose une analyse non conforme à la philosophie de l'ouvrage et, ce faisant, induit le risque d'une lecture erronée du document.

Une deuxième opération, étroitement liée à la première, consiste à *décider jusqu'où sera pris en compte le contenu implicite du document*. Un lecteur peu attentif aura sans doute tendance à comprendre le document selon sa structure de surface, tandis qu'un linguiste métalexicographe ira plus loin dans l'analyse en tentant de mettre au jour sa structure profonde. L'infrastructure du *FEW* a ainsi donné lieu à deux interprétations différentes d'un même type d'information. Dans l'exemple suivant,

> Afr. *bastun* m. „arme de poing" (Roland ; WaceRouA), *baston* (CourLouis—Cotgr 1611 ; Lac […])

l'étiquette géolinguistique « Afr. » qui précède la liste des formes *bastun* et *baston* était comprise dans le premier balisage comme hiérarchiquement supérieure à ces formes et aux autres informations, tandis que É. Büchi (1996 : 116-117) la place sur le même pied et explique son absence devant la forme *baston* par un principe d'économie. Cette analyse permet de rétablir l'unité minimale de traitement comme élément essentiel de la structure du *FEW*. Les deux lectures sont valables, mais laquelle choisir ? La manière dont sont insérées les balises joue ici un rôle important. Un balisage XML automatisé, qui se sert d'indicateurs typographiques, se fondera de préférence sur la structure de surface du document, tandis qu'un balisage manuel permettra plus facilement l'insertion d'informations supplémentaires derrière chaque élément du texte. L'automatisation du processus aura souvent pour conséquence la perte provisoire d'une partie de ce qui, dans le document, relève de l'implicite. Si certaines données pourront être rétablies par la suite grâce à des algorithmes spécifiques, d'autres nécessiteront un traitement manuel ; ainsi, les informations suffixales non explicites, mais présentes en structure profonde dans le *FEW*, ne pourront être prises en compte et rétablies que manuellement, après la rétroconversion du dictionnaire.

4.1.2. *Normalisation*

Des analyses divergentes proviennent également du fait que nous tentons de structurer un document déjà rédigé, dont toutes les règles n'ont pas été explicitées au moment de la rédaction. Il s'agit donc de *revenir à la genèse du document* et de comprendre après coup selon quelle logique il a été conçu. Or, un document présente souvent des incohérences, surtout lorsqu'il a été rédigé sur une longue période et par plusieurs mains. Dans le *FEW*, le problème se pose notamment en ce qui concerne le rôle exact des références bibliographiques et les conditions de leur apparition. Les rédacteurs, ne disposant d'aucune consigne claire à cet égard, ont apparemment suivi des pratiques différentes. Le *FEW* comporte pour chaque variété linguistique (ancien français, occitan, wallon, etc.) une liste de sources « canoniques » implicites, consignées dans un répertoire annexe et numérotées selon leur importance. Alors que leur statut de source canonique les rend facultatives dans le programme lexicographique, elles apparaissent de façon irrégulière et avec une signification variable. Certains rédacteurs semblent avoir noté une référence de façon systématique

lorsqu'elle n'était pas la première dans la liste canonique, d'autres uniquement lorsqu'elle apportait une précision supplémentaire ou encore lorsque le statut de la source était particulier. Il s'ensuit que le dictionnaire présente à cet égard une diversité de pratiques dont il faut tenir compte, avec un risque constant de mauvaise interprétation pour les articles rédigés par des auteurs décédés, qu'on ne peut interroger sur leur méthode.

La constatation de cette diversité des pratiques rédactionnelles se heurte à la nécessité de *normaliser la structure du document*. Un trop grand nombre d'irrégularités est à même de nuire à l'informatisation de documents volumineux, car un balisage automatique nécessite l'existence de règles fiables. Lors de l'informatisation du *TLF* par exemple, les rubriques étymologiques n'ont pas été balisées, car jugées trop irrégulières :

> [l]'histoire de l'élaboration du TLF s'étant étalée sur une période de plus de trente ans, nous avons constaté que seule la partie synchronique correspondait à des normes de rédaction relativement stables. Les normes de rédaction de la seconde partie ne se sont stabilisées que fort tardivement, aux environs du tome 11. Avant ce tome, les différentes rubriques sont constituées d'un discours totalement informel qu'il serait vain de vouloir structurer. (J. Dendien & J.-M. Pierrel 2003 : 15)

La grande souplesse microstructurelle dont fait preuve le *FEW* pose évidemment un problème à ce niveau. En ce sens, la première version du balisage était trop rigide pour correspondre à toutes les situations permises par le *FEW*. Par ailleurs, le discours atypique de ce dictionnaire étymologique en rend l'analyse peu aisée selon les standards de balisage proposés par la *Text Encoding Initiative*[11]. Naviguer entre respect des particularismes et normalisation représente donc, dans certains cas, un véritable casse-tête.

Un document tel que le *FEW* peut donc, dans le processus d'informatisation, souffrir soit d'une mauvaise analyse (relevé incomplet et structuration erronée des types d'informations ; perte de la structure profonde), soit d'une mauvaise compréhension des logiques rédactionnelles et d'une normalisation abusive. Ces « dangers » peuvent être partiellement évités en prenant le temps de cerner les spécificités et les caractéristiques intrinsèques du document à informatiser. Il s'agit notamment de penser à tout ce que l'utilisateur final voudra y trouver et, si besoin, de modéliser non seulement sa structure de surface, mais aussi

11 Cf. http://www.tei-c.org.

sa structure profonde. Il s'agit également de revenir à la genèse du document et de veiller à respecter ses particularismes. Le résultat de cette phase d'informatisation dépendra finalement de la rigueur d'analyse et de l'ingéniosité de l'humain qui définira le balisage ; il consistera souvent en un savant équilibre entre ce que contient le document, ce que veulent y trouver les lecteurs et ce que permet l'automatisation du processus. Dans tous les cas, le document aura été analysé et balisé selon un point de vue, après une série de choix : il aura été interprété, avec les risques que comporte un tel traitement.

4.2. Exploitation du document informatisé

Le document une fois balisé, son exploitation est riche de promesses, à la fois pour les éditeurs et pour les utilisateurs. Il devient en effet une base éditoriale et un « multidictionnaire » :

> [l]es avantages de la *rétroconversion* d'un dictionnaire imprimé en *dictionnaire informatisé* sont considérables. En effet, la transformation du contenu d'un dictionnaire en base de données relationnelle permet d'accéder à chaque élément constitutif du texte, la structure logique de celui-ci ayant été rendue explicite. Non seulement le dictionnaire informatisé représente au sens propre le « *multidictionnaire* » par les innombrables types de lectures-consultations qu'il rend possibles, mais c'est une *base éditoriale* qui facilite les corrections, les mises à jour, les extractions, etc. […] (B. Quemada 1991 : 19-20).

4.2.1. *Une base éditoriale*

Un des grands intérêts de l'informatisation est sans aucun doute la facilité de réédition du dictionnaire, voire de mise à jour systématique. Cette fonctionnalité en fait un dictionnaire évolutif. L'accès par Internet permet par ailleurs de recueillir les suggestions de corrections des utilisateurs, ce qui, en démultipliant le nombre de relecteurs, assure une réactualisation quasi permanente. Ce qui semblait un rêve il y a vingt-cinq ans semble donc devenu réalité :

> […] on peut toujours rêver à un dictionnaire total à la Jorge-Luis Borges (« La bibliothèque de Babel »), inlassablement augmenté, remanié, élagué et reclassé par des machines infaillibles, et que tout un chacun pourrait non seulement consulter, mais enrichir de ses trouvailles en tripotant, de chez lui, quelques boutons. (K. Gebhardt 1982 : 33)

Le *FEW* a fait l'objet d'un nombre impressionnant de corrections et d'ajouts, qui se trouvent actuellement relégués dans sa parastructure (articles de revue, monographies, comptes rendus, etc.). L'intégration de ces matériaux

dans le texte même du dictionnaire est un des objectifs de l'informatisation. Néanmoins, lors de l'enquête réalisée en 1996 sur le sujet, certains spécialistes du *FEW* ont réagi à cette idée de façon péremptoire : « le pire, c'est l'idée d'un dictionnaire évolutif pour le *FEW* » (P. Renders en préparation : Annexes). Pourquoi ? Deux « dangers » sont ici à prendre en compte. Le premier concerne le statut même du document. En tant que document historique, œuvre d'une époque, d'un homme et d'une idéologie, le *FEW* doit conserver son intégrité ; œuvre de référence, citée après toute étymologisation d'un lexème galloroman, il doit pouvoir à tout moment être consulté dans l'état qui était le sien au moment de la citation. En devenant évolutif, il pourrait ne plus être considéré comme un document historique. En revanche, en se tenant à la pointe du progrès, il peut renforcer son statut d'œuvre de référence, à condition toutefois que les modifications qui y sont apportées respectent l'œuvre et sa valeur. Dans l'encyclopédie électronique *Encarta*, par exemple,

> [l]es modalités de cette réactualisation sont cependant strictement définies : une encyclopédie « sérieuse » ne peut en effet ouvrir l'accès qu'à des sites qu'elle a elle-même validés afin de préserver l'homogénéité des données et la fiabilité des informations. (Ch. Jacquet-Pfau 2005 : 60)

Il est donc nécessaire que les mises à jour soient, d'une part, validées par des experts scientifiques et, d'autre part, bien distinguées comme telles avec trace de la date et de l'auteur de la modification. Sans contrôle éditorial et sans journalisation, un *FEW* évolutif trahirait l'œuvre originale, qui ne serait plus consultée dans sa version imprimée, mais ne serait pas non plus transmise adéquatement dans la version électronique.

Un second danger, moins évident à écarter, guette le dictionnaire évolutif : celui de la déstructuration du tissu textuel. Dans le *FEW*, la séquentialité des données est hautement significative ; de plus, le principe d'économie qui sous-tend les règles de rédaction permet la mise en facteur commun de certaines informations[12]. Modifier, ajouter ou supprimer une donnée, quelle qu'elle soit, est susceptible de briser la séquentialité du discours et de provoquer la perte d'informations importantes pour d'autres unités qui en dépendaient. Il est donc important de réfléchir avec soin à la façon dont les mises à jour peuvent être intégrées dans le document sans l'endommager.

12 Cf. É. Büchi (1996 : 117).

4.2.2. *Un multidictionnaire*

L'apport le plus remarquable de l'informatisation est sans aucun doute le changement de pratique auquel elle incite : « [i]l va de soi […] que la lecture-consultation ou exploration des données va prendre le pas sur la lecture linéaire et aboutir à une modification des habitudes des usagers de dictionnaires » (G. Gorcy 1990 : 204). En ce qui concerne le *FEW*, c'est plutôt l'inverse qui se passe, puisque les utilisateurs font appel à l'informatique dans le but de faciliter ces pratiques déjà courantes, mais actuellement ardues. La consultation du *FEW* se résume rarement à la recherche d'un étymon ou d'un lexème ; de nombreux linguistes, notamment les rédacteurs d'autres dictionnaires ou atlas linguistiques comme l'*ALW*, se servent de ce thesaurus pour confronter leurs hypothèses aux données de toute la Galloromania, interrogeant le *FEW* à partir de questions très générales comme l'évolution de tel suffixe latin ou les désignations de tel concept[13]. Ces requêtes sont des exemples types de « lectures transversales » qui parcourent la totalité du texte dictionnairique pour y sélectionner certaines données selon des critères précis.

L'informatisation rendrait donc possible ce type de consultation non linéaire. Or, n'en déplaise aux utilisateurs du *FEW*, ce dernier n'a pas été conçu dans une telle optique ; au contraire, sa macro- comme sa microstructure ont été réfléchies de manière à ce que le classement des lexèmes constitue en lui-même un élément important de leur analyse (cf. § 3). Nous nous trouvons donc devant un paradoxe : les difficultés de consultation du *FEW* nécessitent son informatisation, mais s'il existe tant de difficultés, c'est justement parce que le *FEW* est conçu pour être lu de façon linéaire, chaque article étant à appréhender comme une monographie. L'informatisation permettrait un mode de lecture du document qui irait à l'encontre de la philosophie et de la conception de celui-ci.

Prenons l'exemple d'une requête visant à extraire du dictionnaire tous les lexèmes désignant le cresson, afin de répertorier, de localiser et d'étymologiser ses diverses dénominations dans les différentes régions de l'espace « galloromanophone ». Il est tout d'abord évident que le résultat ne sera pas exhaustif, ne serait-ce que parce que le *FEW* ne relie pas de façon explicite chaque lexème à la notion qu'il désigne. Le seul moyen serait de repérer les

13 Cf. É. Büchi (1992).

définitions qui contiennent le mot *cresson*, ce qui produira inévitablement du bruit (des définitions seront reprises, où le mot a une autre fonction que celle de synonyme du lexème : *cressiculteur* « celui qui cultive le cresson » dans *FEW* 16, 385a) et des lacunes (des définitions seront oubliées, qui définissent un lexème désignant le référent, sans pour autant citer le mot *cresson* : apr. *creisson* « nasturtium officinale » dans *FEW* 16, 384b). Ce genre de recherche nécessiterait donc un traitement préalable du document, par exemple sous forme d'index onomasiologique[14].

Imaginons à présent que l'utilisateur ait réussi à extraire tous les lexèmes désignant le cresson. Que peut-il faire avec le résultat de cette extraction ? Beaucoup de choses et, en même temps, très peu. Beaucoup, car il est facile, à partir d'une liste de mots retournée par l'ordinateur, de se construire rapidement une idée de la question, de réorganiser les données par période, de les représenter sous forme de tableaux statistiques et, par exemple, de dessiner une carte spatiale des désignations du concept. Toutefois, sans retour au texte « linéaire », la plupart de ces opérations ne résisteront pas à un examen scrupuleux et se révéleront partiellement inexactes. L'index du *FEW* nous mettait déjà en garde :

> Une bonne utilisation du *FEW* comporte une lecture (au moins cursive) de l'article complet dont relève la lexie à laquelle on s'intéresse, afin de la situer dans son contexte diasystématique (historique, dialectal, etc.). Nous invitons donc les lecteurs à ne jamais se contenter de localiser la lexie recherchée dans une page précise, sous peine de passer à côté d'informations primordiales (ATILF 2003 : VIII).

Sans une lecture attentive du *FEW*, nous ne pourrions pas deviner que le languedocien *coudèrlo*, pourtant défini « esp. de cresson » (*FEW* 21, 122b cresson), désigne surtout un champignon (*FEW* 21, 165a champignon), ni que les lexèmes relevés dans le Sud de la France tels que le languedocien *graissous* ou le béarnais *crechoû* sont, malgré leur classement sous l'étymon *KRESSO (*FEW* 16, 384b), des emprunts au français qui ont été influencés par la famille du latin CRESCERE. L'extraction d'un lexème non replacé dans son contexte provoque par conséquent la perte de ces informations adjacentes.

L'informatisation d'un document comme le *FEW* amènerait donc chez les utilisateurs un changement de pratique qui leur rendrait d'énormes services

14 L'établissement d'un index onomasiologique de tous les lexèmes du *FEW* est en cours de réalisation à l'ATILF.

en termes de recherche, mais qui, par sa superficialité, provoquerait la perte d'informations essentielles s'ils ne prennent pas le temps de lire attentivement les articles auxquels renvoie leur requête et, donc, de revenir à une pratique plus classique du document. Or, les facilités nouvelles permises par l'informatisation risquent de provoquer ce genre de consultation superficielle, d'autant plus lorsqu'il s'agit de répondre rapidement à une question apparemment simple (« quel est l'étymon du languedocien *graissous* ? »).

Il n'est bien sûr pas question de renoncer à une évolution qui va dans le sens des souhaits des utilisateurs et permet des recherches auparavant impossibles. Il est toutefois nécessaire de réduire l'hiatus entre la conception linéaire originelle du document et le mode de lecture non linéaire favorisé par l'informatique. Une des solutions consiste à prévenir le lecteur, qui doit savoir

1) quelles informations et quelles méthodes de recherche sont valables et lesquelles ne le sont pas, la distinction entre les deux ensembles étant assurée par les principes de la critique historique ou, dans le cas d'un dictionnaire comme le *FEW*, par la métalexicographie ;
2) quelles démarches supplémentaires lui incombent pour assurer la pertinence de sa recherche. Dans le cas d'une consultation transversale, le lecteur doit être d'autant plus formé à la lecture linéaire du *FEW* ; il doit pouvoir retourner au document original et nuancer le résultat de sa recherche par une analyse correcte du contexte dans lequel se trouvent les données extraites.

Un *FEW* informatisé devrait donc remédier lui-même aux défauts de son informatisation et guider l'utilisateur en lui suggérant les « bonnes » manières de procéder. Le traitement préalable du document peut évidemment jouer un rôle dans cette « prévention des risques », une autre part du travail étant assurée par la bonne définition des requêtes possibles et par une conception adéquate de l'interface d'utilisation du dictionnaire. Si, lors du traitement préalable du document, il était essentiel de comprendre comment il avait été conçu, l'exploitation du document informatisé demande que l'utilisateur sache en outre comment il a été traité lors de l'informatisation elle-même. Le *TLFi* par exemple ne permet une utilisation vraiment efficace — et des pratiques différentes par rapport au dictionnaire imprimé — que si l'on sait comment il a été balisé.

5. Conclusion

L'informatisation d'un document n'est pas un acte banal, et encore moins une solution miracle. Nous pourrions l'assimiler au processus difficile de la traduction. Tous deux ont pour objectif premier de rendre l'œuvre plus accessible et de l'ouvrir à un public plus large. Tous deux tentent, ce faisant, de rester fidèles au document original, qui, sauf catastrophe, reste de toute façon conservé et donc accessible dans son état primitif. Tous deux comportent néanmoins des risques, à la fois dans le traitement du document et dans l'exploitation qui sera faite du résultat.

Le traducteur est susceptible de mal comprendre le texte, de ne pas relever certains procédés (un jeu de mots, une figure de style non explicite ou impossible à traduire, etc.), de devoir choisir entre plusieurs traductions possibles, dont aucune n'est totalement satisfaisante. Il peut effectuer une traduction « littérale », qui reste fidèle à la lettre du texte, ou plus « littéraire », en modifiant par exemple la syntaxe pour rendre toutes les nuances stylistiques et autres dans la langue cible. L'informatisation est confrontée aux mêmes types de problèmes. Un balisage peut s'en tenir à la structure de surface du document ou intégrer une analyse plus approfondie de celui-ci, qui en retire toute la substance. Dans les deux cas, le risque existe d'oublier ou d'interpréter erronément certaines données. Un « bon » balisage dépend donc de la finesse d'analyse du « traducteur » (ou « passeur »), de sa connaissance de l'œuvre, de sa capacité à déceler l'implicite et à le rendre au mieux dans la version électronique, afin que le lecteur qui utilise uniquement la version « traduite » ait accès à un maximum d'informations.

Le lecteur, quant à lui, a sa part de responsabilité dans l'exploitation qu'il fera du document traduit ou informatisé. Selon ce qu'il cherche dans le document, il peut être nécessaire de retourner à l'original, sans quoi nombre d'informations seront perdues, tandis que d'autres se révéleront inexactes — ce qui porterait atteinte à la définition même du document (« pièce écrite, servant d'information ou de preuve » selon le *TLFi*).

Comme la traduction, l'informatisation est donc une opération utile et essentielle, mais elle demande que l'on reste conscient des difficultés et des risques qu'elle comporte. Est-ce une coïncidence si *informatisation* évoque à la fois *information* — à ne pas perdre dans le processus — et *formation* — à assurer à l'utilisateur ?

Bibliographie

ALW = L. Remacle, E. Legros, J. Lechanteur, M.-Th. Counet & M.-G. Boutier (1953–), *Atlas linguistique de la Wallonie*, Liège.

ATILF (2003), *Französisches Etymologisches Wörterbuch, Index*, publié par ATILF-CNRS sous la direction d'Éva Büchi, 2 vol., Paris.

K. Baldinger (1974), *Le FEW de Walther von Wartburg. Introduction*, dans K. Baldinger [dir.] (1974), *Introduction aux dictionnaires les plus importants pour l'histoire du français*, *Bulletin des Jeunes Romanistes* 18-19 (= Paris, Klincksieck), p. 11-47.

——— (1980). *Étymologies doubles dans le FEW*, dans *Italic and Romance: Linguistic Studies in Honor of Ernst Pulgram*, Amsterdam, p. 189-194.

Beiheft = W. von Wartburg (1950[2] [1929[1]]), *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes. Beiheft: Ortsnamenregister, Literaturverzeichnis, Übersichtskarte*, Tübingen.

Beiheft Supplement = M. Hoffert (1989 [1957]), *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes. Supplement zur 2. Auflage des Bibliographischen Beiheftes*, Bâle.

É. Büchi (1992), *Le traitement des déonomastiques dans le FEW*, dans G. Hilty [éd.] (1992), *Actes du XX[e] Congrès International de Linguistique et Philologie Romanes*, vol. IV, Tübingen, p. 69-78.

——— (1996), *Les Structures du* Französisches Etymologisches Wörterbuch. *Recherches métalexicographiques et métalexicologiques*, Tübingen.

——— (2003), *Introduction*, dans *ATILF*, Paris, p. V-X.

J. Dendien & J.-M. Pierrel (2003), *Le Trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence*, dans *Traitement automatique des langues* 43/2, p. 11-37.

FEW = W. von Wartburg *et al.* (1922-2002), *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes*, 25 vol., Bonn – Heidelberg – Leipzig/Berlin – Bâle.

K. Gebhardt (1982), *L'apport des dialectes d'oïl (surtout entre 1300 et 1600) au lexique de la langue commune (d'après le FEW)*, dans P. Wunderli [éd.] (1982), *Du mot au texte. Actes du III[e] Colloque International sur le Moyen Français*, Tübingen, p. 31-48.

G. Gorcy (1990), *Le* Trésor de la langue française *(*TLF*). Son originalité et les voies ouvertes par son informatisation*, dans *INaLF*, *Autour d'un dictionnaire : le* Trésor de la langue française, Paris, p. 187-207.

Ch. Jacquet-Pfau (2005), *Pour un nouveau dictionnaire informatisé*, dans *Éla* 137, p. 51-71.

R. Martin (1999), *Perspectives en lexicographie informatisée. L'expérience du* DMF *(*Dictionnaire du Moyen Français*)*, dans *Société de linguistique de Paris*, *Lexique, lexicologie, lexicographie*, Louvain, p. 51-71.

B. Quemada (1991), *Acquis et perspectives de l'informatique*, dans *Travaux de linguistique* 23, p. 17-21.

P. Renders (à paraître), *L'informatisation du* Französisches Etymologisches Wörterbuch *: quels objectifs, quelles possibilités ?*, dans *Actes du XXV^e^ Congrès International de Linguistique et de Philologie Romanes* (Innsbruck, 3-8 septembre 2007), Tübingen.

——— (en préparation), *Modélisation d'un discours étymologique. Prolégomènes à l'informatisation du* FEW, Liège, Université de Liège [Thèse de doctorat].

A. Rey (1971), *Le dictionnaire étymologique de W. von Wartburg : structures d'une description diachronique du lexique*, dans *Langue française* 10, p. 83-106.

G.-M. de Schryver (2003), *Lexicographers' Dreams in the Electronic-Dictionary Age*, dans *International Journal of Lexicography* 16/2, p. 143-199.

TLF = P. Imbs [dir.] (1971–1994), *Trésor de la langue française. Dictionnaire de la langue du XIX^e^ et du XX^e^ siècle (1789-1960)*, 16 vol., Paris.

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