L'informatisation du FEW: attentes et modélisation

Pascale Renders
Université de Liège & Centre National de la Recherche Scientifique et Université de Nancy

Christelle Nissille
Université de Nancy et Université de Neuchâtel

The computerisation project of the Französisches Etymologisches Wörterbuch *aims at taking this reference dictionary on Romance linguistics out of its current state of under-utilization resulting from the complexity of its structures. In view of this, we have submitted on September 2007 a questionnaire designed to better understand the wishes and common practices of FEW users. In the present paper, we examine the first results of the survey and its impact on the future electronic modelisation of the FEW by focusing on cross-searching fields. Implicit information, such as dates, regionalisms or suffixes cannot be found automatically, except by using external tools, that must therefore be taken into account in the computerisation. The solutions considered here do not mean that "classic reading" will become obsolete. Nevertheless, we hope that we will be able to give users new ways to get to the dictionary thus allowing a new and more efficient use of the FEW...*

1. Introduction

Ouvrage de référence en lexicologie et lexicographie historique française et romane, le *Französisches Etymologisches Wörterbuch* (FEW) de Walther von Wartburg présente, dans une perspective historique et étymologique, l'ensemble du lexique des quatre langues galloromanes (français, francoprovençal, occitan et gascon) et de tous leurs dialectes, depuis les premières attestations au 9^{e} siècle jusqu'à l'époque contemporaine. Les matériaux qu'il répertorie sont accessibles dans leur intégralité (de manière exhaustive), classés et hiérarchisés de manière raisonnée selon leurs affinités (génétiques, géohistoriques, sémantiques, phonétiques et morphosyntaxiques).

Cette visée ambitieuse a deux conséquences directes: l'ampleur de l'ouvrage—il s'étend sur 25 volumes et 16.700 pages—et la complexité de ses structures, étudiées par Büchi 1996. De plus, il s'agit d'une œuvre en mouvement, puisqu'elle s'est développée dans un souci constant à la fois d'amélioration et de tradition, que ce soit par Wartburg lui-même ou par les différentes équipes de rédaction qui lui ont succédé. Actuellement, le Centre du FEW, situé à Nancy (ATILF, CNRS/Nancy Université), se consacre à la refonte des articles du premier volume (*cf. www.atilf.fr/few*). Cette entreprise a notamment donné lieu à une réflexion théorique sur les pratiques lexicographiques propres au FEW et à une remise en question de certaines d'entre elles (analyse philologique plus fine, meilleure utilisation des atlas, prise en compte plus importante des informations morphosyntaxiques et sémantiques; *cf.* Chambon 1989b, Chauveau 2005, 2006). Le FEW est devenu ainsi, et dans la lignée de la volonté de son fondateur, à la fois plus philologique et plus linguistique (*cf.* Chambon 1989a; 1989b).

Ce perfectionnement constant s'accompagne d'un souci de rendre le FEW plus accessible, souci qui s'est concrétisé par l'amélioration de la péristructure (mise à jour et informatisation du répertoire bibliographique qu'est le *Beiheft*, réalisation d'un index des formes, etc.) ainsi que par d'autres idées novatrices (*cf.* Chauveau 2006b), parmi lesquelles le désir ambitieux d'informatiser l'ensemble de l'ouvrage. Le projet qui s'est mis en place présente deux faces complémentaires, combinant production (rédaction modulaire en format XML, cf. Matthey / Nissille à paraître) et consultation, par "rétroconversion" du dictionnaire papier en dictionnaire électronique (*cf.* Renders à paraître).

L'informatisation du FEW se heurte à deux contraintes. La première est inhérente à l'objet FEW, puisque sa complexité, qui est à l'origine du projet d'informatisation, est en même temps un obstacle à ce dernier. La deuxième contrainte est la philosophie qui sous-tend ce projet. Comme la réflexion théorique qui l'accompagne, il est au service des utilisateurs du FEW, que ceux-ci soient rédacteurs ou lecteurs. La prise en compte des attentes des utilisateurs rend évidemment la tâche plus ardue: aux obstacles techniques dus à la complexité structurelle du FEW s'ajoutent des *desiderata* parfois peu réalisables en pratique. Dans le cadre d'une étude de faisabilité, il nous semble toutefois important de mener une réflexion en profondeur et d'envisager tout ce qui peut contribuer à ce que le produit fini (le dictionnaire informatisé) réponde au maximum aux besoins des bénéficiaires. Cela permet d'éviter l'écueil consistant à adopter des solutions déjà existantes pour d'autres ouvrages lexicographiques qui, si elles sont à n'en pas douter adaptées aux supports pour lesquelles elles ont été développées, peuvent toutefois limiter les possibilités de traitement d'un ouvrage aussi particulier que l'est le FEW.

Dans cet article, après analyse des demandes formulées par les utilisateurs, nous nous proposons d'étudier leur impact sur la modélisation du discours lexicographique. Ce faisant, notre réflexion se focalisera sur une problématique précise, à savoir le souhait exprimé par les utilisateurs de pouvoir effectuer des recherches transversales.

2. Attentes des utilisateurs

Un sondage auprès des utilisateurs a été effectué de trois manières: par la lecture de publications qui se réfèrent au FEW, par des interviews et discussions informelles et, enfin, par le dépouillement d'un questionnaire diffusé à cet effet. Ce dernier avait pour objectifs, d'une part d'obtenir un aperçu des pratiques des utilisateurs et de déterminer quels types d'informations les intéressent généralement, d'autre part de connaître leurs souhaits, même les plus ambitieux.

Le projet d'informatisation du FEW hérite d'une longue tradition: il est en effet rêvé depuis longtemps par les linguistes (Wooldridge 1990: 239). Plusieurs publications font état des problèmes rencontrés dans la consultation de l'ouvrage. Outre des difficultés d'accès pratique et de lecture du contenu (Rey 1971: 103-104), on relève un repérage malaisé de certains types d'informations (par exemple les déonomastiques, Büchi 1996: 69-70) et, de manière plus générale, le manque de fiabilité des résultats de recherche, qui ne peuvent jamais être tenues pour exhaustives (Baldinger 1974: 25). Les auteurs regrettent aussi l'impossibilité d'effectuer une mise à jour des données, ou encore de pouvoir extraire un ensemble d'unités d'une synchronie donnée (Wooldridge 1998: 211).

Les réponses au questionnaire confirment ces relevés et permettent de les préciser. On constate ainsi des demandes unanimes:[1]

- du point de vue de la lecture: résoudre par simple "clic" les nombreuses *abréviations* contenues dans le dictionnaire (sigles géohistoriques et bibliographiques);
- du point de vue de la recherche: questionner le dictionnaire *de façon transversale* pour y relever un ensemble de mots ayant une ou plusieurs caractéristiques particulières;
- du point de vue du dialogue avec d'autres données: bénéficier de *liens hypertextuels* avec les données fournies par d'autres dictionnaires informatisés;
- du point de vue de la mise à jour: accéder à une *mise à jour des données*, qu'il s'agisse d'ajouts et corrections minimes (datation, source, localisation, etc.) ou plus importantes (étymologisations nouvelles et réécriture de parties d'article).

Parmi ces souhaits, celui portant sur les recherches transversales est particulièrement intéressant et retiendra notre attention dans la suite de l'exposé. Il s'agit, concrètement, de sélectionner

[1] Nous ne ne listons pas ici les attentes plus ponctuelles et moins partagées.

dans le dictionnaire une série d'unités lexicales répondant à un ou plusieurs critères bien définis, tels que[2]:

- une même appartenance linguistique (en l'occurrence dialectale): "tout le lexique de l'ancien dauphinois";
- une même catégorie grammaticale;
- une même propriété morphologique: "tous les composés N-N", "tous les mots formés à l'aide d'un certain affixe";
- un même signifié: "tous les lexèmes francoprovençaux désignant le cresson";
- une même datation: "tous les proverbes du 16e siècle";
- un ensemble d'étymons présentant un même suffixe: "tous les mots qui proviennent d'un étymon en -ACULUM", "tous les mots dont les étymons se terminent par -ATEM mais non -CATEM";
- les étymons de même langue: "tous les mots normands provenant du norrois".

Comme le montrent ces exemples, les critères peuvent évidemment se combiner. Une recherche sur "tous les adjectifs occitans munis du suffixe *-enc*", associe ainsi informations géolinguistique, grammaticale et morphostructurelle.

Dans la structure du discours lexicographique du FEW, ces critères ne sont pas sur le même pied. Certains sont en effet facilement repérables dans le dictionnaire (appartenance géolinguistique, étymon) tandis que d'autres sont considérés par les utilisateurs comme difficilement accessibles (propriétés sémantiques ou morphologiques). Dans le point suivant, nous nous proposons d'analyser le statut de ces informations dans le discours fewien et de voir ce qu'implique, pour la modélisation du FEW, la prise en compte de ces critères de recherche.

3. Impact sur la modélisation du FEW

3.1. *Problématique*

Rendre le FEW exploitable pour toute sorte d'interrogations suppose d'avoir préalablement acquis le texte sous forme électronique, ce dont nous ne traiterons pas ici. Il s'agit ensuite d'y insérer un balisage XML adéquat, qui reflète la structure du texte et qui permette les requêtes souhaitées. La conception de ce balisage est un des objectifs de l'étape de modélisation. Elle revêt une importance capitale, puisque c'est à ce moment que sont —ou ne sont pas— prises en compte les attentes des utilisateurs. Les balises devant ensuite être insérées dans le texte de façon le plus possible automatisée, il faudrait en outre que chaque type d'informations qui fait l'objet d'un balisage puisse être reconnu au moyen d'indicateurs formels fiables (*cf*. Renders à paraître; Dendien et Pierrel 2003).

L'attitude la plus conventionnelle consisterait à définir le balisage uniquement en fonction des capacités de reconnaissance des informations par l'automate, en laissant de côté les aspects de la structure dont l'automatisation pose problème. Si cette démarche se justifie pour certains textes, où une telle restriction ne porte pas préjudice à la qualité des informations obtenues, un accès trop restreint constituerait pour le FEW un handicap majeur, puisque c'est dans la complémentarité de toutes les données que réside l'intérêt de l'œuvre de Wartburg. Il est d'autant plus nécessaire de la prendre en compte que c'est cette somme d'informations que cherchent à atteindre les utilisateurs.

Lorsque l'on tente de modéliser la structure du FEW, un problème se pose immédiatement. Büchi (1996: 117) montre que l'ouvrage possède, derrière sa structure de surface, une structure profonde qui s'en écarte à plusieurs niveaux. Cette différence naît du principe d'économie suivie par le dictionnaire et de l'implicite que génère ce principe. Il a pour but d'éviter de

[2] Les exemples fournis proviennent des réponses au questionnaire.

répéter des informations afin de gagner de la place —argument d'importance pour un dictionnaire papier— et de gagner en lisibilité par synthétisation des informations.

C'est évidemment la structure profonde que doivent pouvoir atteindre les utilisateurs lors des recherches transversales. Elle doit donc être prise en compte dans la modélisation, même si son inaccessibilité partielle en surface pose un problème dans la conception des algorithmes de reconnaissance. Le principe de ces algorithmes étant justement qu'ils se basent sur la reconnaissance des formes, la principale question est de savoir comment leur permettre de repérer les informations cachées, qu'elles soient partiellement accessibles ou totalement inaccessibles.

3.2. *Analyse*

3.2.1. *Informations de la structure profonde qui sont partiellement accessibles*

Par *informations partiellement accessibles* nous entendons des informations qui n'apparaissent pas directement comme telles, mais sont cachées sous d'autres éléments de la structure de surface. Ce procédé est mis en œuvre dans l'infrastructure du discours fewien, qui se présente comme une suite d'unités minimales de traitement, elles-mêmes constituées de "molécules" (Büchi 1996: 116). Ces molécules sont au nombre de huit et apparaissent comme suit:

> Étiquette géolinguistique, *signifiant*, catégorie grammaticale, "signifié" (informations complémentaires, localisation, datation, référence bibliographique)

Les quatre premières molécules constituent un cas particulier. Elles sont obligatoires, mais ne sont pas répétées, selon le principe d'économie suivi par le FEW, lorsqu'elles ont déjà été citées auparavant. Elles sont donc facilement restituables par simple lecture des unités qui précèdent. Dans l'exemple suivant, l'étiquette géolinguistique, la catégorie grammaticale et la définition de *phantasie* et de *fantaisie* sont à reprendre de la première unité *fantasie:*

> Fr. *fantasie* f. "imagination reproductrice ou créatrice" (12. jh.–Montaigne), *phantasie* (Or 1370; MistSQ; Calvin), *fantaisie* (Cohen Rég–Ac 1932; "vieux" Voltaire Dict. phil.; "style didact." Trév 1771–Lar 1932; "c'est du didactique suranné" Fér 1787). (FEW 8, 360a, PHANTASIA)

Les quatre autres molécules (informations complémentaires, localisation, datation et référence bibliographique), qui sont souvent indiquées entre parenthèses, sont facultatives. Elles sont toutefois souvent présentes au moins en structure profonde. La référence bibliographique, par exemple, peut ne pas être mentionnée et est alors sous-entendue par l'étiquette géolinguistique qui précède la forme:

> norm. *fontaisie* "fantaisie" DT, nant. Blain *fantasie*, St-Seurin, SeudreS. *fontaisie*, morv. *fantasie*, Moselle f ã t ä Î y Ö y , Cleurie, St-Nab. *fantauhie*, Aussois, Ruff. f ã t a z i , Chav. *fantejio;* nfr. *phantasie* "inclination; caprice amoureux" (BPériers; Rab), *fantaisie* (Régnier–Delv 1867); *fantasie* "caprice en gén., bizarrerie" D'Aubigné, *fantaisie* (seit Cotgr 1611). (FEW 8, 360a, PHANTASIA)

Dans cet extrait, alors que pour le lexème normand, une information explicite sur la source est donnée ("DT"), ce n'est pas le cas de ceux catégorisés par "nant.", "Moselle", "Cleurie", etc. Ces unités lexicales sont néanmoins attestées dans des sources canoniques qui traitent des parlers de ces régions et qui sont implicitement comprises sous chaque étiquette.

Le même procédé est à l'œuvre pour la *date*. Alors qu'elle est explicitement indiquée dans certains cas ("Delv 1867"), dans d'autres elle est implicitement comprise dans les références bibliographiques données après la forme: "BPériers", "Rab", "D'Aubigné" renvoient à des sources datées. Une date peut même se cacher derrière une localisation. Büchi (1996: 126) précise en effet qu'au niveau de la structure profonde, la molécule de la datation est toujours implicitement présente, puisque les localisations renvoient à des sources, qui sont, elles, datées […]".

Büchi poursuit en affirmant que "tout le matériel n'est donc pas daté, mais tout le matériel est datable à l'aide du *Beiheft*". Le système de référenciation du FEW fonctionne donc en symbiose

avec le *Beiheft* et le *Beiheft Supplement*, répertoires bibliographiques qui permettent de résoudre les abréviations, de situer les toponymes cités, d'obtenir diverses informations critiques sur les sources et d'avoir accès, pour chaque étiquette géolinguistique, à la bibliographie des sources qui y correspondent. Dans notre exemple, l'étiquette "nant." et les sigles "BPériers", "Rab", "Régnier" ou "D'Aubigné" font l'objet des explicitations suivantes:

> **localisation et référence bibliographique:**
>
> nant. = nantais, mundart von Nantes (Loire-Inf.): 1. Eudel, P., Les locutions nantaises; N. 1884. — 2. Gaumer, V., Répertoire du langage populaire du pays nantais (Manuskript im besitz der stadtbibl. zu Nantes. Die materialien beziehen sich auf die populäre sprechweise im gebiet der unteren Loire. Sehr reichhaltiges glossar, beinahe 8000 wörter, das auch eine große sammlung von redensarten enthält). — 3. Locutions et prononciations vicieuses, usitées à Nantes et dans plusieurs autres villes occidentales de la France; Nantes s. d. (wird mit 1820 gekennzeichnet). (*Beiheft* et *Beiheft Supplement*)
>
> datation:
>
> BPériers: Frank, F., et Chenevière, A., Lexique de la langue de Bonaventure des Périers; Paris 1888 (BPérier starb ca. 1544. Auch Despériers abgekürzt);
>
> Rab: Rabelais (Das 1. buch ist von 1534, das 2. von 1532, das 3. von 1546, das 4. von 1552, das 5. von 1564, doch nur zum teil von Rabelais, der 1553 gestorben ist);
>
> Régnier: Les Fortunes et Adversitez de Jean Régnier; p. p. E. Droz; SAT; Paris 1923 (der dichter stammt aus Auxerre, lebte ca. 1392 bis 1468; die dichtung ist nach 1433 entstanden);
>
> D'Aubigné: Œuvres complètes de Th. Agrippa d'Aubigné, t. 6e, glossaire; Paris 1892 (D'A. lebte 1560—1630). (*Beiheft* et *Beiheft Supplement*)

Le FEW peut dès lors se permettre d'être très laconique en structure de surface, le rédacteur pouvant généralement choisir de donner une localisation, une datation ou une abréviation bibliographique selon ce qui lui semble le plus économique ou le plus pertinent, chacune de ces trois informations renvoyant aux autres via le *Beiheft*.

On mesure donc l'erreur qu'il y aurait, pour une recherche transversale automatisée, à rester à un niveau de surface: comment répondre à une requête portant sur "les mots liégeois datant du XVIIe siècle" si l'automate ne prend pas en compte pour la datation autant les dates explicites que les sources (dans ce cas par exemple "Haust"[3]) correspondant à la période envisagée? Ainsi, au niveau de la structure informatique qui devra être mise en place, le *Beiheft* joue un rôle central. Il est non seulement la solution au souhait exprimé par les utilisateurs de pouvoir résoudre les abréviations utilisées dans le dictionnaire, mais également la clé pour retrouver les informations implicites.

Une modélisation efficace, prenant en compte la structure profonde du FEW, doit dès lors s'articuler avec une informatisation intelligente du *Beiheft*. Il s'agit tout d'abord de déterminer avec précision et dans le détail les modalités d'apparition de l'implicite dans le discours fewien, en revenant à la genèse du discours et aux choix opérés lors de la rédaction des articles. Il est ensuite possible, dans la version informatisée du *Beiheft*, de mettre en évidence, sous chaque abréviation, les informations pertinentes. Le lien avec le dictionnaire est assuré par un balisage du texte fewien qui donne accès à cet outil: chaque donnée contenue dans un article peut ainsi être reliée aux localisations, dates et sources éventuelles qui lui correspondent.

Le *Beiheft* une fois modélisé et transformé de la sorte, il sera possible de s'en servir pour effectuer de façon automatisée une série de recherches transversales. Reprenons par exemple les *desiderata* de Wooldridge (1998: 211) concernant le vocabulaire du français au 16e siècle:

> "La seule façon de mettre au jour ce qui concerne le français du XVIe siècle dans le FEW serait d'informatiser les 25 volumes... puis de les interroger à partir de repères comme "fr.", "mfr.", "16e s.", etc."

[3] Haust = Haust, J., Le dialecte liégeois au XVIIe siècle; Bibl.de la Faculté de philosophie et lettres de l'Université de Liège, fasc. XXVIII; Liège 1921. (*Beiheft*)

Ce type d'interrogations, tel que Wooldridge l'envisage, peut se contenter d'une lecture en structure de surface, à condition toutefois que l'utilisateur connaisse les notations (abréviations, dates, etc.) du FEW pertinentes dans le cadre de sa recherche. En revanche, si l'on veut affiner les résultats et rechercher par exemple le vocabulaire qui *apparaît* au XVI^{e} siècle, il faut ajouter aux repères mentionnés par Wooldridge (étiquette géolinguistique et mentions explicites d'une datation) les sigles de documentation qui sont datés du XVI^{e} siècle (par exemple "Andernacht", "Aneau", "ArtNav"), dont l'apparition dans le corps des articles indique en règle générale une première attestation. On trouverait ainsi, toujours dans l'article PHANTASIA, les formes référencées par les sigles "BPériers", "Rab" et "D'Aubigné", datés du XVI^{e} siècle dans le *Beiheft*:

> nfr. *phantasie* "inclination; caprice amoureux" (BPériers; Rab) (...) *fantasie* "caprice en gén., bizarrerie" D'Aubigné (FEW 8, 360a, PHANTASIA)

Un autre exemple de recherche, souhaitée par Lagueunière 1998, qui est rendue possible par le biais d'une informatisation du *Beiheft* concerne le caractère régional des lexèmes. En effet, si cette information est explicitement donnée dans les articles de la refonte, ce n'est pas le cas dans les volumes précédents, Wartburg n'ayant pas fait la différence entre forme dialectale et forme régionale:

> "Le problème majeur réside, en fait, dans le traitement des variantes géographiques du français contemporain. Ces matériaux, peuvent être très riches dans le *FEW*, pour certaines familles de mots, mais ils sont difficiles à détacher d'un ensemble de données présenté globalement comme dialectal." (Lagueunière 1998: 387)

Il est uniquement possible de repérer ces informations via les commentaires critiques contenus dans le *Beiheft*, qui indiquent si une source comprend des régionalismes. Par exemple, sous l'abréviation "nant." se trouve la source suivante:

> nant.: 3. Locutions et prononciations vicieuses, usitées à Nantes et dans plusieurs autres villes occidentales de la France; Nantes s. d. (wird mit 1820 gekennzeichnet). (*Beiheft*)

L'expression "nant. *porée* (schon 1820)" que l'on trouve dans le FEW (9, 194b) renvoie à cette source, comme l'indique l'étiquette géolinguistique associée à la date citée entre parenthèses. Dans ce cas, la modélisation du *Beiheft* doit mettre en évidence à la fois la date et la caractéristique régionale de la source, afin qu'une lecture automatique du FEW qui rencontrerait pour une même unité l'étiquette "nant." et la date "1820" puisse en déduire l'information "régionalisme". Il faut donc prendre en compte non seulement l'information géolinguistique, mais aussi les indications qui permettent de caractériser la source précise dont il est question.

3.2.2. *Informations de la structure profonde qui sont inaccessibles*

Les types d'informations traités précédemment étaient certes implicites, mais restaient accessibles, car ils étaient régis par la structure au sens large (comprenant l'épistructure) du FEW. Il faut à présent traiter le cas de l'inaccessibilité de certains éléments de la structure profonde du dictionnaire. Il s'agit d'informations utilisées comme principe organisateur des matériaux, mais qui ne sont jamais explicitées (si ce n'est parfois dans la refonte). Nous pouvons signaler comme relevant de ce cas les éléments formants des lexèmes:

> II. 1. a. Afr. *plurel* m. "pluriel" GuernesSThomas; adj. „qui marque la pluralité" (hap. 13. jh.). — Ablt. Afr. *pluralment* "ensemble" ChGuill, *plurellement* adv. "au pluriel" (hap. 13. jh.), *plurelment* (hap. 13. jh.), nfr. *plurelment* (ca. 1380, Aalma 9365).
>
> b. Afr. *plurer* m. "pluriel" (GuernesSThomas, variante; Gillon), mfr. id. (ca. 1430), apr. *nombre pluzar* (Manosque 1293). (FEW 9, 101a, PLURALIS)

Dans l'extrait ci-dessus, les suffixes servent de critère de regroupement des lexèmes. Toutefois, comme aucune marque explicite ne les caractérise en dehors du commentaire final de l'article PLURALIS, une requête automatique portant sur ces éléments formants s'avère impossible. Nous pouvons néanmoins essayer de prendre en compte ce souhait des utilisateurs en élargissant la perspective, comme nous l'avons fait dans le point précédent avec le *Beiheft*.

Prenons l'exemple d'une requête visant à repérer tous les mots de l'ancien occitan formés avec le suffixe -IOLUM. On aimerait avoir accès à une table de toutes les formes indexées, où seraient

indiquées leur formation et la mention de leur suffixe éventuel. Malheureusement, un tel répertoire n'existe pas encore et son élaboration, qui ne pourrait se faire que manuellement, serait trop fastidieuse. De quoi disposons-nous pour l'instant, qui pourrait apporter une aide? Nous avons, depuis 2003, l'index raisonné du FEW (ATILF 2003) reprenant un peu plus de 275.000 unités accompagnées de leur adresse dans le dictionnaire, de leur étymon et d'une étiquette géolinguistique normalisée. Il est possible, en nous servant de cet outil et de listes de suffixes déjà établies (notamment au Centre du FEW), de rechercher dans l'index tous les étymons se terminant par un suffixe précis (dans notre cas -IOLUM), de repérer les finales des attestations correspondant à ces étymons, puis d'en faire une liste par étiquette géolinguistique, tout cela de façon automatisée.[4] Cela permettrait une ébauche de correspondance entre le suffixe latin et les finales dialectales, qui peut être vérifiée à l'aide d'œuvres traitant des affixes (par exemple Nyrop 1904 ou Adams 1913, qui permettent de distinguer les formations suffixales régulières des réfections).

Cette correspondance peut alors servir d'outil pour la recherche dans le dictionnaire, par un double critère: à la fois par l'étiquette géolinguistique sélectionnée (dans notre cas "apr.") et par une reconnaissance formelle de la finale correspondante. Afin d'éviter le bruit (formes présentant la même finale mais ne provenant pas de ce suffixe), il est possible de prendre en compte le contexte dans lequel l'unité lexicale apparaît: il s'agit de vérifier dans son entourage si d'autres unités (appartenant à d'autres étiquettes géolinguistiques) présentent une correspondance attestant la même construction.

Un survol rapide de l'index permet dans notre exemple de repérer les étymons ANCILLARIOLUS, ARIOLUS, AULAEOLUM, BALNEOLUM, BASIOLUM, BRACHIOLUM, CAPREOLUS, MEDIOLUM, MŎDIOLUS, MORTARIOLUM, OSTIOLUM, *ROSARIOLUS et VITREOLUS. Les lexèmes cités dans l'index pour ces étymons comportent les finales suivantes:

> afr.: *-uel;* mfr.: *-eoux, -eux, -eul -ole;* frm.: *-eul, -eu, -ole, -iol;* awall.: *-uel;* aflandr.: *-eul;* apic.: *-ou, -ieu, -uel;* apr.: *-ol;* arouerg.: *-ol;* agasc.: *-ol;* wall.: *-ol, -ou;* norm.: *-ole*, *-eu*,; hbret.: *-eul;* manc. *-œ;* orl.: *-ieux;* lorr.: *-euil;* Bresse: *-ole;* dauph. (frpr.): -·; dauph. (occit.): -µw; pr.: *-òu;* lang.: *-ol*, *-oul*, -µ*l;* rouerg.: -∑*l*, *-ou, -uól;* lim.: *-ol;* gasc.: *-oou, -òu,* -∑l, *-ò, -ó*·.

Ce résultat[5] permet de chercher dans le FEW les lexèmes de l'ancien occitan comportant la finale *-ol*. On trouverait ainsi, sous PARIUM (7, 656a), les items suivants:

> Apr. *pairol* m. "chaudron" (seit 13. jh., Rn; MeyerDoc; Gdf 5, 696; AnnAlpes 5, 278), *peyrol* (14. jh., R 34, 193).

La pertinence de ce résultat est confirmée par le contexte, puisque le reste du paragraphe comporte des formes correspondantes, notamment des formes de l'occitan en *-ou:* pr. *peir*∏·, mars. *peiroou* A, Aix *peirou* P, etc[6].

3.3. *Solution: une troisième dimension*

Les exemples précédents montrent que, pour permettre aux utilisateurs d'effectuer les recherches transversales exprimées dans les réponses aux questionnaires, il s'avère nécessaire de modéliser le FEW en tenant compte de sa structure profonde et en donnant accès à tout ce

---

[4] On peut éventuellement envisager ici un traitement manuel intermédiaire, consistant à éloigner les formes doublement suffixées comme *-iolittu*.

[5] Pour que le résultat soit tout à fait correct, il est évidemment nécessaire de trier les formes non pertinentes. En ce qui concerne par exemple MEDIOLUM, il a fallu enlever les formes présentes sous le paragraphe "suffw" et les formations tardives (sv. 2: avec suff. -ONE).

[6] Outre le fait de repérer dans le FEW le résultat de recherches ponctuelles (ici les formes de l'ancien provençal en -IOLUM), il est possible de se servir de cette procédure pour constituer de mini-corpus. Dans l'exemple traité ici, il s'agirait de sélectionner tous les paragraphes du FEW regroupant les formes selon le critère du suffixe -IOLUM afin de les analyser. Ce corpus pourrait servir dans une recherche visant, par exemple, à distinguer les changements suffixaux (voir à ce propos l'article BĀSIOLUM, *www.atilf.fr/few*).

qui, dans le discours lexicographique fewien, est de l'ordre de l'implicite. Nous avons distingué deux sortes d'implicite: un implicite accessible car faisant l'objet d'une explicitation dans un autre endroit du FEW (notamment dans le *Beiheft:* datations et sources) et un implicite à priori inaccessible car non résolu de façon systématique ailleurs dans le discours fewien (affixes).

Dans les deux cas précités, une solution partielle a pu être dégagée grâce à une démarche semblable. Elle consiste à "sortir du cadre" et à accéder à une troisième dimension via l'utilisation d'outils annexes: le *Beiheft* dans le premier cas, l'index dans le second. La modélisation du FEW doit donc prendre en compte ces outils et les articuler avec le discours contenu dans les articles du dictionnaire. Cette démarche assimile le texte fewien à une "parole" au sens saussurien du terme, le *Beiheft* et les index jouant le rôle de dictionnaires de la langue "few", puisqu'ils font le relevé des mots autorisés et expliquent entre autres leur sémantisme individuel et les rapports sémantiques entre eux. Une modélisation du FEW consiste dès lors à annoter chaque mot du texte en y apportant les informations "sémantiques" qui conviennent et sur lesquelles pourront porter les recherches. Comme dans une langue, où les dictionnaires permettent de cerner le sens des mots en contexte et jouent un rôle prépondérant dans le traitement informatique, la parole fewienne bénéficie d'outils de description aptes à aider au décodage et, moyennant quelques aménagements, à permettre le traitement automatique du texte.

4. Conclusion

La prise en compte des attentes des utilisateurs a induit une démarche consistant à analyser la faisabilité de ces demandes et à étudier leur impact sur la modélisation du dictionnaire. Cette étude permet de décider quels souhaits peuvent être rencontrés, lesquels doivent être abandonnés et, en fin de compte, de "remettre une offre". Cette dernière se révèle plus nuancée que prévu. Les exemples traités ci-dessus montrent en effet que les réponses ne sont pas de l'ordre du "oui" ou du "non". Nous répondons plus souvent par un "oui, mais" qui représente une offre partielle: non pas soumise à condition, mais partiellement réalisée, ce qui est très différent. Dans le cas de recherches transversales portant sur une datation ou un suffixe, l'emploi du *Beiheft* et des index permet certes de construire un corpus plus restreint que les 25 volumes du FEW, mais la recherche n'est pas totalement satisfaisante, le résultat fourni pouvant contenir du bruit ou s'avérer non exhaustif.

Alors que la demande des utilisateurs nécessitait d'analyser son impact sur la *modélisation* du FEW, la dernière étape consistera à étudier l'impact de notre offre sur l'*utilisation* du FEW. En effet, puisque les solutions que nous pouvons fournir ne répondent pas toujours de façon totalement fiable aux demandes des utilisateurs, ces derniers vont être mis à contribution pour améliorer le résultat de leur recherche: la restreindre à un niveau inférieur, éliminer le bruit et — toujours!— vérifier les informations fournies. Cette démarche évite par ailleurs les dangers d'aplatissement des données que peut provoquer une consultation informatique "naïve".

Puisque le retour au texte s'impose même après une recherche automatisée —il s'agit non seulement de vérifier la pertinence du résultat fourni par la machine, mais aussi d'accéder à l'analyse linguistique fouillée qui constitue le vrai contenu du FEW—, l'utilisation du dictionnaire informatisé ne devrait donc pas se révéler, finalement, très différente de celle du dictionnaire papier. Ce constat est heureux. Le FEW est en effet un dictionnaire très particulier, un répertoire exhaustif, raisonné et critique, qui ne peut être résumé à une suite de données. Le classement des matériaux qu'il opère reflète une analyse approfondie de ceux-ci, analyse qui ne peut être atteinte de façon mécanique, mais uniquement par une lecture attentive des articles. Buchi (*in* ATILF 2003: 1: VII) prévenait déjà, dans la préface à l'index des formes du FEW, qu'il ne remplace pas "la lecture (au moins cursive) de l'article complet". La consultation du FEW exige une participation du lecteur, qui doit obligatoirement interpréter les données. Quel est, alors, l'intérêt d'une informatisation? Il réside avant tout dans l'accès plus rapide, et meilleur, aux données recherchées, notamment par des chemins auparavant impossibles ou peu praticables. Dans cette optique, le sondage effectué auprès des utilisateurs prend tout son sens. Il est essentiel, dans les limites évoquées ci-dessus, de prendre en compte les voies d'entrée dans le FEW qu'ils privilégient.

Nous espérons avoir démontré l'importance, pour la réussite d'un tel projet, de l'étape de modélisation, seule opération proprement linguistique. Le modèle élaboré en fin de compte devant à la fois refléter le plus justement possible la structure du discours lexicographique étudié et permettre un maximum d'interrogations qui intéressent les utilisateurs tout en restant réalisable d'un point de vue pratique (automatisation du processus), le résultat consiste, dans la plupart des cas, en un compromis entre ces diverses exigences. C'est dans cet équilibre que réside, selon nous, le succès de l'entreprise qui a pour but, en répondant au défi de l'informatisation du FEW, de fournir une meilleure accessibilité au contenu d'"un des plus beaux monuments des sciences du langage" (Swiggers 1990: 347).

Références bibliographiques

ATILF (2003). *Französisches Etymologisches Wörterbuch, Index*. Publié par ATILF-CNRS sous la direction d'Éva Buchi (2 vol.). Paris: Champion.

Adams, E. L. (1913). *Word-formation in Provençal.* New York: The Macmillan Company.

[Beiheft]. Wartburg, W. von. (1950^2 [1929^1]). Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes. Beiheft: Ortsnamenregister, Literaturverzeichnis, Übersichtskarte. Tübingen: Mohr.

[Beiheft Supplement]. Hoffert, M. (1989^2 [1957^1]). Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes. Supplement zur 2. Auflage des Bibliographischen Beiheftes. Bâle: Zbinden.

Büchi, E. (1996). *Les Structures du* Französisches Etymologisches Wörterbuch*. Recherches métalexicographiques et métalexicologiques*. Tübingen: Niemeyer.

Chambon, J. P. (1989a). "Aspects philologiques et linguistiques dans la refonte du FEW: utilité d'une approche métaphilologique des représentations linguistiques". Dans Kremer, D. (éd.). *Actes du 18e Congrès International de Linguistique et Philologie Romanes*. Tübingen: Niemeyer. Vol. 7. 218-230.

Chambon, J. P. (1989b). "Tradition et innovations dans la refonte du FEW". Dans Kremer, D. (éd.). *Actes du 18e Congrès International de Linguistique et Philologie Romanes*. Tübingen: Niemeyer. Vol. 7. 327-337.

Chauveau, J. P. (2005). "Remarques sur la dérivation dans les notices historiques et étymologiques du Trésor de la langue française". Dans Buchi, É. (éd.). *Actes du Séminaire de méthodologie en étymologie et histoire du lexique* (Nancy / ATILF, année universitaire 2005 / 2006). Nancy: ATILF (CNRS / Université Nancy 2 / UHP), publication électronique (*http://www.atilf.fr/atilf/seminaires/Seminaire_Chauveau_2005-11.pdf*).

Chauveau, J. P. (2006). "Sur l'étymologie de fr. *baie* "petit golfe"". *Revue de Linguistique Romane* 70. 409-427.

Chauveau, J. P. (2006b). "D'un site informatique en chantier pour le FEW". Dans Schweickard, W. (éd.). *Nuovi media e lessicografia storica: atti del colloquio in occasione del settantisimo compleano di Max Pfister*. Tübingen: Niemeyer. 33-37.

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

[FEW]. Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes. Bonn/Heidelberg/Leipzig - Berlin/Bâle: Klopp /Winter/Teubner/Zbinden, 1922-2002.

Laguenière, F. (1998). "Le traitement de la variation diatopique en français moderne dans le *Französisches Etymologisches Wörterbuch"*. Dans Ruffino, G. (ed.). *Atti del XXI Congresso Internazionale di Linguistica e Filologia Romanza* (Palermo 18-24 settembre 1995). Tübingen: Niemeyer. 387-395.

Matthey, A. C.; Nissille, C. (à paraître). “L’irruption de l’informatique dans la rédaction du FEW”. Dans *Actes du XXV^e Congrès International de Linguistique et de Philologie Romanes* (Innsbruck, 3-8 septembre 2007). Tübingen: Niemeyer.

Nyrop, K. (1908). *Grammaire historique de la langue française, tome 3. Formation des mots*. Copenhague: Nordisk Forlag.

Renders, P. (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: Niemeyer.

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

Wooldridge, T. R. (1990). “Le FEW et les deux millions de mots d’Estienne-Nicot: deux visages du lexique français”. *Travaux de linguistique et de philologie* 28. 239-316.

Wooldridge, T. R. (1998). “Le lexique français du XVI^e siècle dans le GDFL et le FEW”. *Zeitschrift für romanische Philologie* 114. 210-257.