Vers un enrichissement raisonné de la rétroconversion du *Französisches Etymologisches Wörterbuch* (FEW)

Nicolas Mazziotta, Université de Stuttgart, Institut für Linguistik/Romanistik
Pascale Renders, Université de Liège, Linguistique française et dialectologie wallonne

L'informatisation par rétroconversion du Französisches Etymologisches Wörterbuch *(FEW) de Walther von Wartburg, oeuvre fondamentale de la lexicologie historique galloromane, est à présent en cours de réalisation et nous voudrions montrer en quoi elle est perfectible et comment elle pourrait être améliorée. Nous présentons brièvement le FEW et la nécessité de le rendre exploitable par l'ordinateur (attentes des utilisateurs et besoins de formalisation), puis nous exposons les principes qui ont gouverné sa rétroconversion (au format XML) et donnons l'exemple détaillé de l'article* substantivus, *dont nous faisons la lecture suivie. Ensuite, nous focalisons l'exposé sur la microstructure rétroconvertie des articles et l'enrichissement par annotation manuelle que nous préconisons. Nous montrons quelles informations ne sont pas accessibles à la machine en raison de l'omniprésence de conventions implicites dans l'œuvre originale. Nous synthétiserons enfin le potentiel de notre approche : l'accès à l'*intelligence *du FEW, et non plus seulement à sa* forme.

1. Introduction

L'informatisation par rétroconversion du *Französisches Etymologisches Wörterbuch* (FEW) est en bonne voie et nous voudrions montrer en quoi elle est perfectible et comment elle pourrait être améliorée. Nous présenterons brièvement le FEW et la nécessité de le rendre exploitable par l'ordinateur (→1), puis nous exposerons les principes qui ont gouverné sa rétroconversion et en donnerons un exemple détaillé (→2). Ensuite, nous focaliserons l'exposé sur la microstructure rétroconvertie des articles et l'enrichissement par annotation manuelle que nous préconisons (→3). Nous synthétiserons enfin le potentiel de cette dernière (→4).

2. Le FEW et la nécessité de son informatisation

Le FEW de Walther von Wartburg est une œuvre monumentale (Buchi et Chambon 1995). Il présente, dans une perspective historique et étymologique, l'ensemble du vocabulaire galloroman (français, francoprovençal, occitan et gascon et tous leurs dialectes), depuis les premières attestations (9^{e} s.) jusqu'à l'époque contemporaine. Mais loin d'être un recueil de listes de formes, ce gigantesque thesaurus (25 vol., 16.700 p.) peut être vu comme un immense recueil de monographies: chaque article contient un champ documentaire, où les données sont analysées et classées selon une *structure raisonnée*, propre à chaque article selon les particularités de la famille lexicale traitée. À ce champ s'ajoute parfois un commentaire expliquant le classement. L'ouvrage combine ainsi deux dimensions : thesaurus et ensemble de monographies (voir Büchi et Chambon 1995: 952). Dans la suite de l'exposé, nous utiliserons le terme d'*inscription papier* pour désigner la forme imprimée de cet ouvrage[1].

La complexité du FEW rend son exploitation difficile. Pour remédier à cela, et surtout depuis l'arrivée d'Internet et en particulier la mise en ligne du *Trésor de la langue française* (TLFi), la communauté scientifique demande son informatisation, c'est-à-dire la constitution de ce que nous nommerons une *inscription numérique* de la connaissance qu'il véhicule. Les attentes[2] concernent autant la recherche de données que l'aide à leur lecture. Au niveau de la

[1] Sur l'utilisation du terme *inscription*, voir Bachimont 2007: 10; voir également notre développement dans le cadre qui nous occupe dans Mazziotta, à paraître: § 1.1.

[2] Selon les résultats d'un sondage effectué auprès de 30 utilisateurs du FEW, parmi lesquels on peut compter les

recherche, on voudrait pouvoir effectuer des requêtes simples (p. ex.: toutes les occurrences d'une information spécifique, selon un critère géographique, géolinguistique, chronologique, diasystématique, morphologique, sémantique, etc.) ou des requêtes complexes, combinant plusieurs critères. En ce qui concerne la *lecture*, on voudrait que soient résolus les abréviations et les sigles[3], mais aussi et surtout obtenir facilement le plan des articles longs.

3. Rétroconversion de l'ouvrage

Dans la perspective d'une informatisation, deux voies différentes sont envisageables (voir Mazziotta, à paraître: § 1.2): la conversion directe de l'inscription papier dans une inscription numérique, processus appelé *rétroconversion*, ou la construction d'une nouvelle inscription numérique: une refonte complète dans un environnement qui les rendrait directement exploitables par les machines (Matthey et Nissile, à paraître). Nous ne traiterons ici que de la rétroconversion[4].

3.1. Démarche

La démarche en est la suivante. Tout d'abord, l'inscription papier est numérisée par océrisation ou saisie manuelle (projet subventionné par la DFG) et enrichie immédiatement d'un certain nombre d'éléments XML correspondant à la typographie (p. ex.: les limites des paragraphes, indiquées par l'élément <p>, les formes en italiques, dans l'élément <i>, etc.). En termes d'ingénieurie des connaissances, cette première étape construit une *formalisation des contenus* de l'inscription papier, c'est-à-dire de sa *forme d'expression* (Bachimont 2007: 14-17). Dans un second temps, l'inscription produite est soumise à un automate qui affine la formalisation en cherchant à distinguer des formes apparemment similaires sur la base de critères structurels. Ainsi, une ensemble de chiffres pourra être une date ou une référence à une page en fonction de sa place dans l'article. Cette démarche permet de repérer avec une assez grande fiabilité les structures de l'article ainsi que les différentes *molécules* qui forment l'unité minimale de traitement du discours féwien: étiquette géolinguistique, signifiant, catégorie grammaticale, signifié, précisions complémentaires (nous ne traiterons pas cette molécule ci-dessous)[5]. Toutefois, la deuxième étape n'en demeure pas moins une description de la *forme*.

3.2. Exemple

Lisons ensemble le résultat du passage de l'automate sur la saisie manuelle du champ documentaire de *substantivus* (FEW 12, 357a) — exemple abrégé (les passages omis sont indiqués par <gap>[...]</gap>) et simplifié. L'article commence comme ceci (la plupart des balises sont interprétables intuitivement):

<entry><etymon desc='n/a' type='vedette'> substantivus**</etymon>** für sich selbst bestehend.**</entry>**

---

membres de la Société de Linguistique et de Philologie Romanes. Voir également Renders à paraître.

[3] De nombreuses informations sont abrégées de manière parfois obscure. La consultation du FEW nécessite le recours à plusieurs volumes résolvant les abréviations et les sigles : les *Beihefte* (von Wartburg 1950; Hoffert 1989).

[4] Notez que la lecture de la suite de l'exposé nécessite une compréhesion générale du langage XML (voir *http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html* pour une introduction rapide).

[5] Les molécules citées ici ne sont pas exactement celles proposées initialement par Eva Buchi (1996: 98, 116-129).

Il s'agit de l'entrée de l'article, qui est directement suivie du champ consacré à la documentation:

<doc> <pnum id='1'>1.</pnum> <geoling>Mfr.</geoling> <form>substantif</form> <gram>adj.</gram> <def>„qui forme la base de (p. ex. de la vie religieuse, de l'amour, etc.)'</def> (<date>14. jh.</date>–<biblio>Desch</biblio>, <biblio>Froiss</biblio>, <biblio>Gdf</biblio>; <biblio>Lac</biblio>); <gap>[...]</gap>

Ce qui se lit: '*Substantif* est attesté en moyen français, comme adjectif, au sens de 'qui forme la base de'. Il est attesté chez Deschamps et Froissard d'après les dictionnaires de Godefroy et Lacurne de Sainte Palaye [etc.].'. Poursuivons la lecture:

<pnum id='2 a'>2. a.</pnum> <geoling>Mfr.</geoling> <geoling>nfr.</geoling> mot, nom <form>substantif</form> <def>„qui, sans le secours d'un autre, désigne l'être, la chose qui est l'objet de la pensée'</def> (seit <date>14. jh.</date>), <form>substantif</form> <gram>m.</gram> <def>„mot qui seul, et sans le secours d'un autre, désigne l'être, la chose qui est l'objet de notre pensée'</def> (hap. <date>14. jh.</date>; <date>1550</date>, <biblio>Meigret</biblio>; seit <biblio>Pom 1671</biblio>), <form>sustantif</form> <biblio>Desgr 1821</biblio>; <form>substantif</form> <gram>adj.</gram> <def>„qui appartient au substantif'</def> (<biblio>Fur 1690</biblio>–<biblio>Lar 1875</biblio>; <biblio>Rob</biblio>).

On voit qu'un autre groupe de formes (2.a) peut être délimité. *Substantif*, attesté depuis le moyen français, est un adjectif qui signifie 'qui, sans le secours...'. D'autres mots sont regroupés à cette forme : le nom masculin correspondant, ainsi que l'adjectif qui en découle. À la suite de ces formes, on trouve:

Ablt. — <geoling>Nfr.</geoling> <form>substantivement</form> <gram>adv.</gram> <def>„en manière de substantif'</def> (seit <date>1660</date>); <form>substantival</form> <gram>adj.</gram> <def>„qui a la même nature qu'un substantif'</def> (seit <biblio>Lar 1923</biblio>). — <geoling>Mfr.</geoling> <form>substantiver</form> <gram>v. a.</gram> <def>„faire un substantif (d'un adj., d'un verbe)'</def> (<date>1380</date>, <biblio>Aalma 11938</biblio>; <biblio>Garb 1487</biblio>; <biblio>DuBell</biblio>), <geoling>nfr.</geoling> id. (seit <biblio>AcC 1842</biblio>), <geoling>apr.</geoling> substantivar (hap.); <geoling>nfr.</geoling> <form>substantifier</form> (seit <date>1647</date>; ‚vieilli' <biblio>DG</biblio>); <form>substantification</form> <gram>f.</gram> <def>„action de substantiver'</def> <biblio>Rob 1961</biblio>.

l'abréviation *Ablt.* (*Ableitungen*) indique que les formes qui le suivent sont des dérivés. Enfin,

<pnum id='2 b'>b.</pnum> <gap>[...]</gap> <pnum id='3'>3.</pnum> <geoling>Nfr.</geoling> colorant <form>substantif</form> <def>„fixé par les fibres textiles sans l'intervention d'un mordant'</def> (seit <biblio>Besch 1845</biblio>). </doc>

outre un ensemble de formes plus proches de 2 a. que de 1., un autre ensemble de formes se distingue de 1. et de 2. Comme on le voit, la rétroconversion assigne des formes à des ensembles en se fondant sur des critères formels tels que l'enchaînement des paragraphes numérotés, ce qui pourrait permettre de construire une table des matières reprenant simplement ces numéros.

4. Traitement de la microstructure

4.1. Intervention humaine inévitable

Évaluons le résultat de la rétroconversion par rapport aux exigences de la Communauté (→1). En ce qui concerne les recherches dans le dictionnaire, l'automate ne peut s'aventurer au delà de la structure : les critères d'organisation des formes les unes par rapport aux autres (essentiellement morphologiques et sémantiques) ne sont pas reconnus car ils ne sont donnés

qu'implicitement par la structuration de chaque article. Pour ce qui est de l'aide à la lecture, si la rétroconversion résout les abréviations et explicite les sigles, elle ne permet que partiellement la création de tables des matières, livrant une simple hiérarchie dont la motivation n'est pas exprimée. La rétroconversion ne répond donc *pas* aux besoins qui sont en rapport étroit avec la dimension monographique du FEW, qui requiert un niveau de *compréhension de la microstructure* de chaque article.

Or, à la suite de la documentation figure un champ consacré au commentaire, rédigé en langue naturelle et impossible à formaliser complètement:

> Lt. *substantivus*„für sich selbst bestehend' wird, von dingen gesagt, bereits von Tertullian gebraucht (*res substantiva*). In dieser ältern und weiter ausgreifenden bedeutung ist es im 14. jh. vom fr. entlehnt worden (oben 1) [...]. Der grammatiker Priscian (um 500) bezeichnet das verbum *sum* als ‚verbum substantivum'; wenn es für sich allein, ohne adj. oder andere beigaben als prädikat gebraucht wird. Offenbar wurde es in der grammatischen terminologie des Mittelalters ausgedehnt auf wörter, die dinge oder wesen bedeuten. Daraus im 14. jh. entlehnt 2 a. Die verwendung des wortes, wie sie sich bei Priscian findet, wird im 16. jh. dazu entlehnt (b). In der ursprünglichen bed. „auf sich selbst beruhend' wird es im 19. jh. von der technik entlehnt (3).

À la lecture du commentaire, qui décrit la structure du champ documentaire[6], il saute aux yeux que ce qui n'est pas accessible à la machine reste facilement compréhensible par un humain. Ce dernier aura tôt fait de mettre en relation la microstructure et le commentaire, en associant à chaque paragraphe une valeur spécifique. Le commentaire permet aussi de comprendre que le sens grammatical le plus ancien est représenté au travers de la collocation *verbe substantif*, du groupe 2 b. et non 2 a., alors que c'est ce dernier qui reçoit la priorité à cause de son ancienneté.

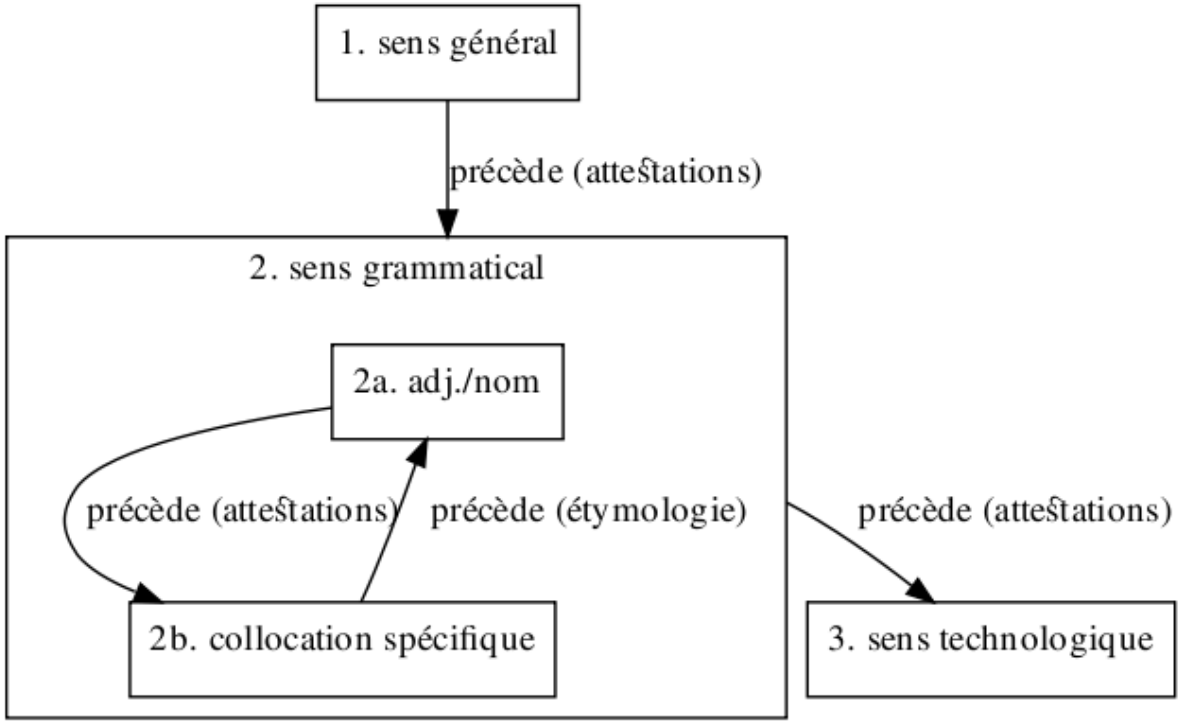
D'autre part, l'humain est également capable de percevoir des relations plus subtiles. Ainsi, si nous reprenons la partie du champ documentaire consacrée aux dérivés, nous voyons que la forme *substantification* est formée sur *substantifier*, ce que l'auteur de l'article a indiqué par la simple succession des formes. Or, cette succession peut avoir bien d'autres valeurs sans qu'un indice structurel ne le montre.

4.2. Application

Il appert donc que, pour répondre aux besoins des utilisateurs, la microstructure rétroconvertie doit être enrichie par une annotation manuelle explicitant: 1/ la nature des regroupements en paragraphes (avec une hiérarchie à plusieurs niveaux); 2/ la nature du lien entre les regroupements en paragraphes[7]; facultativement, 3/ les regroupements non marqués (comme celui qui lie *substantifier* à son déverbal). Ainsi, pour l'exemple qui nous occupe, nous devrions construire (pour les deux premiers points) le graphe suivant:

[6] Ce n'est pas toujours le cas, cf. Büchi 1996: 154-155.

[7] Dans certains cas non abordés ici, comme celui de *personalis* (FEW 8, 273a), où le jeu des emprunts mène à une structuration principale partiellement fondée sur la transmission, les liens sémantiques éventuels sont indiqués dans le commentaire.



Nous n'aborderons pas ici le formalisme que nous préconisons[8]. Pour l'heure, nous nous bornerons à dire que cette annotation opère par regroupement de classes. Nous définissons donc un inventaire de classes très générales dont la portée est l'ensemble du FEW. Par exemple, la classe des *Formes regroupées selon un critère sémantique* (GS) ou celle des *Formes regroupées selon un critère de transmission* (GT) ou *Formes regroupées selon leur classe morphosyntaxique* (GM). Nous spécialisons ensuite ces classes en sous-classes spécifiques à l'article étudié. Ainsi, pour l'exemple qui nous occupe, trois sous-classes distinctes de GS seront identifiées: celle des formes au sens *général*, celle des formes au sens *grammatical* (subdivisée en deux classes) et celle des formes au sens technologique. Les relations entre les classes sont ensuite explicitées. Ainsi, non seulement les relations de précédence représentées par la structure de l'article (ici, le critère d'ancienneté des attestations) sont explicitées, mais celles qui n'apparaissent pas de prime abord (comme le fait que la collocation *verbe substantif* est plus proche du sens étymologique) le sont également.

5. Exploitations

En ajoutant une couche descriptive qui explicite le raisonnement sous-tendant la structure de l'article, cette formalisation enrichit la rétroconversion préalable et permet:

1. d'extraire la structure conceptuelle des articles et de construire une table des matières;
2. de personnaliser la création de cette table des matières en inversant la précédence des critères (par exemple, pour *substantivus*, en plaçant le groupement basé sur GM avant celui basé sur GS);
3. de comparer la structure conceptuelle d'articles traitant de mots de la même famille (non traité ici);
4. de formuler des requêtes du type 'quels articles sont fondés sur une distinction sémantique au premier chef?', ou 'quels articles sont structurés sur la base d'une

[8] À savoir *OWL Ontology Web Language* (utilisation proposée dans Mazziotta, à paraître).

opposition entre les termes grammaticaux et les autres mots?';

En d'autres termes, elle permet d'atteindre le classement des données, leur *analyse*, et dépasse la simple création de listes, basées sur des critères plus ou moins complexes. Elle accède à l'intelligence de l'œuvre.

Références

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

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

Bachimont, B. (2007). *Ingénierie des connaissances et des contenus. Le numérique entre ontologies et documents*. Paris: Lavoisier.

Büchi, E.; Chambon, J.-P. (1995). 'Un des plus beaux monuments des sciences du langage : le FEW de Walther von Wartburg (1910-1940)'. Dans Antoine, G.; Martin, R. *Histoire de la langue française, 1914-1945*. Paris: CNRS Editions. 935-963.

Büchi, E. (1996). *Les structures du Französisches Etymologisches Wörterbuch.* Tübingen: Niemeyer.

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

Matthey, A.-C.; Nissille, C. (à paraître). 'L'irruption de l'informatique dans la rédaction du FEWi'. Dans *Actes du XXVe Congrès International de Linguistique et de Philologie Romanes (Innsbruck, 3-8 septembre 2007)*. Tübingen: Niemeyer.

Mazziotta, N. (à paraître en 2011). 'L'informatisation du Französisches Etymologisches Wörterbuch. Concepts pour une approche modélisée commune à l'Atlas Linguistique de la Wallonie'. Dans *ZrP* 127.

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

TLF = Imbs, P. (dir.). (1971–1994). *Trésor de la langue française. Dictionnaire de la langue du XIXe et du XXe siècle (1789-1960)*. 16 vol. Paris.

TLFi = CNRS/Université Nancy2/ATILF (2004). *Trésor de la Langue Française informatisé* (cédérom). Paris: CNRS Éditions. http://stella.atilf.fr/ [consulté le 15 mars 2010].