Gilles Souvay & Pascale Renders

2.3. Traitement informatique

1 Introduction

Depuis le *Trésor de la langue française* (TLF), de nombreux projets lexicographiques se sont succédé à l'ATILF et ont profité de l'expérience de ceux qui les ont précédés. L'arrivée de l'informatique a, notamment, profondément modifié la manière de travailler des lexicographes. Le *Dictionnaire du Moyen Français* (DMF, cf. DMF2012), dont la première version remonte à une époque où internet n'était pas encore aussi présent qu'aujourd'hui (DMF1 ; cf. Martin 1999), a ainsi montré l'intérêt de publier directement en ligne un dictionnaire scientifique : son accès, sa consultation et sa mise à jour en ont été grandement facilités. À sa suite, d'autres projets lexicographiques menés au sein de l'ATILF ont bénéficié de la vitrine qu'offrait internet et des développements informatiques sous-jacents. Dans ce contexte, il est rapidement apparu utile de dégager les besoins communs, d'appliquer une méthodologie partagée et de développer des outils permettant de rationaliser les développements : de ce constat est née une plate-forme lexicographique qui a été conçue, mise en œuvre et mise à jour par l'un d'entre nous (cf. Martin/Gerner/Souvay 2010) et qui constitue aujourd'hui le support commun de tous les projets lexicographiques de l'équipe « Linguistique historique française et romane » de l'ATILF. Par ailleurs, la réflexion menée autour de l'informatisation du *Französisches Etymologisches Wörterbuch* (FEW, cf. Renders à paraître) a fait apparaître les principaux pièges à éviter dans la conception d'un dictionnaire étymologique électronique, ainsi que la nécessité de prendre en compte, dès l'étape de structuration des articles, l'exploitation qui en sera faite par les utilisateurs.

Le DÉRom s'intègre dans cet environnement lexicographique nancéien et bénéficie des outils développés pour les autres projets. Même si la présente contribution se trouve dans une version imprimée du DÉRom, celui-ci se présente comme un dictionnaire purement électronique, de sa rédaction à sa consultation : il relève donc de la lexicographie informatique, par opposition à la lexicographie informatisée. Nous développons ci-dessous les particularités d'abord de sa rédaction, ensuite de sa consultation sur la plate-forme lexicographique de l'ATILF.

2 Rédaction informatique du DÉRom

2.1 Saisie des articles en XML

La rédaction des articles du DÉRom s'effectue via une formalisation en XML (*Extensible Markup Language*).[1] L'utilisation de ce langage informatique de balisage générique des informations présente au moins trois avantages par rapport à une rédaction via un traitement de texte commercial. Le premier réside dans la pérennité du format XML, qui se présente sous la forme d'un fichier texte non propriétaire. N'étant pas lié à une version donnée du logiciel qui a permis de le créer, un article XML sera toujours éditable : cette propriété garantit, sur des projets lexicographiques de longue durée, que les données antérieures resteront accessibles. Un deuxième avantage provient du fait qu'un article rédigé en XML ne nécessite pas de rétroconversion, c'est-à-dire qu'il n'est pas nécessaire de le modifier pour le rendre compréhensible par un programme informatique (contrairement à ce que l'on constate par exemple pour le processus d'informatisation du TLF ou du FEW, cf. Dendien/Pierrel 2003 et Renders à paraître). Les fichiers XML sont directement exploitables via de nombreux outils standard ; l'informaticien a le loisir également de développer ses propres outils de traitement.

Le langage XML a enfin pour avantage majeur de garantir l'homogénéité de la rédaction. Un article de dictionnaire possède en effet une structure relativement contrainte. Les différents éléments sont hiérarchisés et ordonnés de façon précise ; certains sont obligatoires, d'autres facultatifs. Selon la nature de l'élément, la typographie (taille de la police de caractères, usage des grasses etc.) peut varier. Le lexicographe qui utilise simplement un traitement de texte se voit obligé de garder en mémoire tous ces détails : il doit se focaliser à la fois sur le contenu, sur la structuration interne et sur le rendu typographique de son article. Dans le meilleur des cas, le résultat sera parfait du point de vue du contenu, mais, étant donné les limitations de l'esprit humain, des problèmes portant sur la structure et la typographie sont presque inévitables. Ceci est d'autant plus vrai lorsqu'un nombre important de rédacteurs interviennent dans l'élaboration du dictionnaire. De ce fait, il s'avère nécessaire de procéder *a posteriori* à une révision formelle des articles, afin de garantir leur homogénéité et faciliter leur lecture. Or, l'utilisation du langage XML rend superflu ce processus

1 Pour une introduction rapide à XML, voir http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html.

de révision formelle. Il permet en effet de contraindre fortement la saisie des articles, de façon à imposer le respect des règles de structuration et de présentation déterminées pour un dictionnaire donné. Le rédacteur peut, dès lors, rester focalisé sur le contenu de son article plutôt que sur la structure et la forme.

La saisie d'un article XML s'effectue à l'aide d'un logiciel de balisage, qui y associe un schéma et une feuille de style. Le schéma structure l'article en spécifiant le jeu de balises valides et la logique de leur enchaînement. Il existe des schémas prédéfinis (voir, par exemple, les recommandations de la *Text Encoding Initiative*, http://www.tei-c.org), mais il est également possible de créer un schéma spécifique au projet. On parle de *document valide XML* si le schéma est respecté. Les membres d'un projet sont supposés échanger entre eux des documents valides ; les articles doivent être valides pour leur publication. La feuille de style permet quant à elle de gérer automatiquement la typographie et l'habillage du texte balisé. C'est grâce à elle que, dans le DÉRom, la séquence <signifiant>xxx</signifiant> <signifiant>yyy</signifiant> s'affiche automatiquement *xxx*/*yyy* (mise en italique des signifiants et séparation des deux signifiants par une barre oblique), sans intervention aucune du rédacteur.

2.2 Structuration XML d'un article du DÉRom

Les spécificités du DÉRom ont conduit à développer un schéma XML propre au dictionnaire. Ce schéma est susceptible d'ajustements mineurs, mais sa trame générale ne devrait pas changer de manière significative au cours du projet. La structure du DÉRom a cela de commun avec le FEW qu'on peut y distinguer une microstructure et une infrastructure (cf. Büchi 1996, 5–6). Nous décrivons ci-dessous le balisage XML de ces deux niveaux ; voir également l'« Avis au lecteur » sur le site internet du DÉRom (sous « Consultation du dictionnaire»), ainsi que Buchi/Gouvert/Greub 2014 pour la structuration de la partie documentaire.

Un fichier informatique du DÉRom contient en général un seul article, identifié à l'aide de la balise <Article>. La microstructure d'un article est composée de sept grandes parties, définies chacune par un élément XML. Après l'entrée de l'article (<Lemme>), le DÉRom, comme le FEW, distingue une partie documentaire (<Materiaux>) et une partie de commentaire (<Commentaire>). Suivent ensuite la bibliographie générale (<Bibliographie>), les signatures des différents intervenants (<Signatures>) et les informations de publication (<MiseEnLigne>), qui contiennent la première date de mise en ligne et la date de la version courante. La dernière partie de l'article, facultative, est constituée des notes (<Notes>), dont le rôle consiste à expliciter les appels de note insérés dans les autres parties de l'article.

L'entrée de l'article est elle-même subdivisée en trois parties obligatoires : la forme de l'étymon reconstruit (<Signifiant>), sa catégorie grammaticale (<Catgramm>) et sa définition (<Signifie>). La partie documentaire, consacrée aux matériaux, fait éventuellement l'objet de subdivisions, marquées par la balise <subdiv> ; chaque subdivision est alors explicitée par un marqueur numérique et un titre, tous deux inclus dans l'élément <titre>. Suivent obligatoirement (et directement, en l'absence de subdivisions) l'énoncé du mot-forme qui représente l'étymon direct des cognats regroupés dans le paragraphe (<etym>), ensuite la documentation proprement dite (<cognats>), découpée par unités de base (<cognat>). Le contenu de l'élément XML <cognat>, qui procure des informations sur les unités lexicales formant la base de la reconstruction comparative, constitue l'infrastructure du DÉRom, décrite ci-dessous. Le commentaire est quant à lui constitué de paragraphes (<p>) contenant du texte libre, à l'intérieur duquel des balises peuvent néanmoins identifier des éléments particuliers. Certains de ces éléments XML apparaissent obligatoirement : c'est le cas notamment du signifiant (<etymsignifiant>) et du signifié (<etymsignifie>, balise qui contient à son tour les balises <analytique> [définition componentielle] et <glose> [définition sous forme de glose rapide]) de l'étymon reconstruit. D'autres, comme le corrélat latin de l'étymon reconstruit (<correlatlatin>), sont facultatifs.[2]

L'infrastructure du DÉRom (<cognat>) est composée, en structure profonde, de cinq éléments : l'idiome dont relève le cognat (<idiome>), son signifiant (<signifiant>), sa catégorie grammaticale (<catgramm>), son signifié (<signifie>) et les références bibliographiques assurant son existence (<refbibl>). Les éléments <signifiant> et <refbibl> sont obligatoires, c'est-à-dire qu'ils apparaissent toujours en structure de surface. En revanche, les éléments <catgramm> et <signifie> (exceptionnellement aussi l'élément <idiome>) sont facultatifs, c'est-à-dire qu'ils sont élidés en structure de surface si leur contenu est identique à celui du cognat qui les précède directement. Ces règles rappellent celles qui régissent l'implicite dans l'infrastructure du FEW (cf. Büchi 1996, 117 et Renders à paraître, 76–81).

L'élément <refbibl> est structuré plus finement. Il peut contenir une datation (<datation>), des précisions, notamment au sujet de la première attestation (<precision>), et une succession de sources (marquées chacune par la balise <reference>). Il est remarquable de constater que, dans le DÉRom comme dans le FEW (dont la structuration informatique a pourtant été effectuée *a posteriori*),

2 Le caractère facultatif de la mention du corrélat latin concerne le seul niveau informatique : les normes rédactionnelles en prévoient une mention obligatoire quand il existe.

une limite a dû être établie dans le découpage XML de l'infrastructure. L'élément <date>, par exemple, pourrait se subdiviser en sous-balises permettant d'exprimer la complexité des formats de datation utilisés : *1362*, *dp. av. ca 1362*, *apr. 1362*, *2^e^ m. 14^e^ s.* etc. À l'usage, la saisie d'un article étant déjà relativement complexe, il est apparu qu'un excès d'éléments XML nuisait au confort de saisie des articles en alourdissant inutilement la rédaction. Certains éléments au contenu bien structuré comme la balise <date> ont donc été définis dans le schéma XML comme contenant du texte libre ; c'est une procédure externe à la saisie qui vérifie leur contenu. Dans le cas d'une rédaction informatique tout comme dans celui d'une rétroconversion *a posteriori*, la construction d'un schéma XML reste un compromis entre la volonté de structuration fine (garantissant l'homogénéité de la rédaction) et la nécessité de ménager, malgré tout, liberté et souplesse.

3 Exploitation informatique du DÉRom

3.1 Une plate-forme centralisée

D'un point de vue informatique, la rédaction du DÉRom a pour particularité d'être distribuée, les rédacteurs travaillant de façon autonome localement : les articles sont rédigés avec un outil externe de saisie balisée. En revanche, la publication des articles et leur consultation s'effectue via une plate-forme centralisée. La plate-forme porte le nom d'ISIS, acronyme de *Interrogations Simplifiées et Interfaces Simplifiées de données au format XML*. Elle s'appuie sur un ensemble de travaux réalisés en lexicographie au laboratoire ATILF. Les concepts initiaux ont été développés pour le DMF au début des années 2000, puis généralisés pour le projet TLF-Étym afin d'obtenir un outil paramétrable en fonction du dictionnaire à mettre en ligne. L'idée de départ est de simplifier la tâche de l'administrateur informatique en lui offrant un éventail prédéfini de fonctionnalités disponibles pour chaque nouveau projet. La plate-forme a été initialement écrite en C/C++ avant de migrer dans un environnement PHP/SQL pour suivre l'évolution des techniques informatiques. Cette réécriture doit permettre à terme de la rendre portable et de la distribuer en *open source*.

Le DÉRom constitue une instance parmi d'autres (une petite dizaine actuellement) de la plate-forme et n'en exploite pas encore toutes les fonctionnalités. La plate-forme se compose de quatre parties : une composante lexicographique, une composante textuelle, une composante bibliographique et une composante de gestion des pages du site. La composante lexicographique per-

met de consulter les articles selon des formulaires de recherche prédéfinis, configurés à partir des éléments XML propres au dictionnaire. La composante textuelle n'est pas active pour le projet DÉRom : elle offre la possibilité de gérer un corpus de textes et d'accéder à des attestations qui n'auraient pas été sélectionnées par le rédacteur (voir par exemple DMF2012). La composante bibliographique permet d'expliciter les abréviations des ouvrages cités dans le corps des articles. Enfin, le module de gestion des pages est chargé de gérer l'ensemble des pages du site. Nous décrivons ci-dessous uniquement les fonctionnalités lexicographiques et bibliographiques de la plate-forme qui sont utilisées par le DÉRom.

3.2 Typologie des utilisateurs

La plate-forme distingue cinq types d'utilisateurs : le consultant, le rédacteur, le responsable de la bibliographie, le responsable du projet et l'administrateur informatique. Selon son statut, l'utilisateur a accès à différents éléments et fonctionnalités de chaque composante. Le consultant accède par exemple aux articles publiés, à la bibliographie et aux informations publiques concernant le projet et l'aide en ligne. Le rédacteur accède en outre aux articles privés (articles en cours d'élaboration partagés entre rédacteurs et réviseurs) et aux informations relatives à la documentation. Il dispose de quelques outils en ligne d'aide à la rédaction (contrôle d'articles avant publication) et de fonctionnalités supplémentaires pour la consultation des articles (champs masqués aux consultants). Le responsable de la bibliographie utilise les fonctionnalités permettant de mettre à jour la bibliographie. Enfin, la mise à jour des articles publics et des pages du site est réservée au responsable de projet, tandis que l'administrateur accède à toutes les fonctionnalités de la plate-forme.

3.3 Publication des articles

La plate-forme offre les fonctionnalités classiques de gestion d'une base (ajout d'un nouvel article, suppression d'un article existant, réinitialisation de l'ensemble). Ces fonctionnalités sont uniquement accessibles au responsable du projet, qui valide, lors de la mise en ligne, le contenu scientifique de l'article. Le responsable du projet peut donc poster les articles en ligne sans attendre la disponibilité de l'informaticien : il est autonome et occupe la fonction d'administrateur de la base.

La composante lexicographique de la plate-forme offre par ailleurs aux rédacteurs, avant publication des articles, quelques aides et outils de vérification. Une fonctionnalité dite « de contrôle » permet notamment de vérifier les informations qui ne faisaient pas l'objet de validation lors de la saisie ou qui sont impactées par des modifications dues à l'évolution du schéma XML. L'ensemble des références bibliographiques d'un article du DÉRom, par exemple, sont systématiquement vérifiées : celles absentes de la bibliographie ou ne respectant pas le format de citation sont signalées. Citons encore la fonctionnalité de visualisation, qui affiche l'apparence qu'aura l'article lors de sa mise en ligne, et la fonctionnalité de partage en ligne des articles qui se présentent dans un état avancé, mais qui ne sont pas encore publiables tels quels.

3.4 Consultation du dictionnaire

La consultation des articles publiés bénéficie de fonctionnalités de recherche accessibles au consultant et au rédacteur. Le contenu de chaque élément XML se classe selon un des trois types de données suivants, auquel sont associés des formulaires différents de recherche : la liste d'entrées, la liste de valeurs prédéfinies et le texte libre. L'interface de consultation du DÉRom utilise une partie seulement des formulaires proposés. Les listes d'entrées (les entrées pouvant être des lemmes, mais aussi des noms de rédacteurs, par exemple) y sont essentiellement accessibles via des formulaires du type « nomenclature », qui permettent un affichage paginé avec recherche via l'initiale des entrées (cf. ci-dessous figure 1). La liste de valeurs prédéfinies, qui permet une recherche ciblée d'informations selon des critères linguistiques particuliers, n'est pas encore utilisée pour le projet DÉRom à l'heure où nous écrivons ces lignes (pour un exemple, voir le formulaire de recherche sur les classes étymologiques du projet TLF-Étym). Enfin, le formulaire associé au texte libre permet de rechercher une chaîne de caractères dans un élément XML donné. Actuellement, ce formulaire est utilisé pour la recherche plein texte dans l'ensemble du DÉRom, mais il pourra plus tard être proposé de façon plus ciblée, par exemple pour rechercher une chaîne de caractères dans une glose.

Les éléments du DÉRom proposés à l'interrogation sont divers. Un article du dictionnaire (« Consultation du dictionnaire par articles ») peut être atteint soit via l'étymon protoroman qui constitue le lemme de l'article (<Lemme>), soit via son corrélat latin, soit via l'entrée correspondante du REW$_3$. Des consultations sont également proposées via la recherche d'informations appartenant à l'infrastructure du DÉRom : cognats romans (atteints via l'élément XML <signifiant>), signifiés (balise <glose> à l'intérieur de la balise <etymsignifie>), catégo-

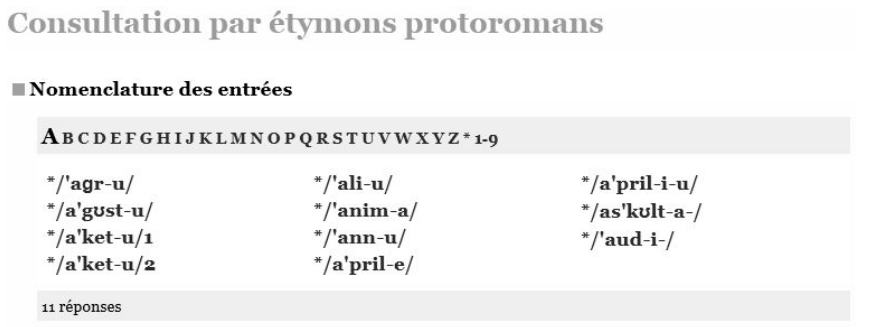

Figure 1 : Formulaire d'interrogation des entrées du DÉRom

ries grammaticales (<catgramm>) ou idiomes romans (<idiome>). La recherche d'une forme protoromane permet d'atteindre les étymons protoromans cités dans le commentaire ou dans les notes. Enfin, l'interface propose également une recherche par collaborateur, distinguant entre rédacteurs, réviseurs et contributeurs ponctuels.

Le résultat d'une requête affiche soit la liste des entrées qui correspondent à la demande, soit les articles eux-mêmes. Par exemple, une consultation du DÉRom par rédacteur affiche d'abord la liste des rédacteurs (figure 2) ; un clic permet ensuite d'accéder à la liste des articles du rédacteur choisi (figure 3), puis aux articles eux-mêmes.

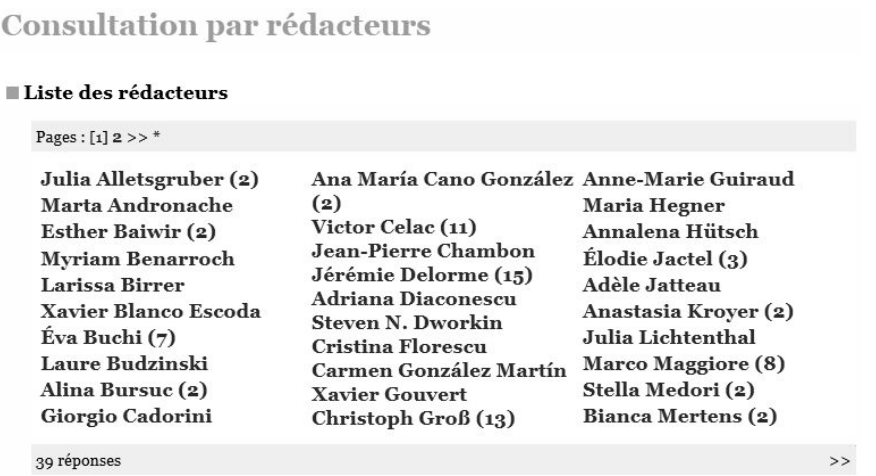

Figure 2 : Liste des rédacteurs du DÉRom

Figure 3 : Liste des articles d'un rédacteur

3.5 Gestion de la bibliographie

La composante bibliographique de la plate-forme permet d'expliciter les abréviations des ouvrages cités dans le corps des articles. Le consultant peut cliquer sur une référence bibliographique pour obtenir son extension complète. Il dispose aussi d'un formulaire de recherche sur l'ensemble de la bibliographie et peut télécharger la bibliographie complète du projet. La responsable de la bibliographie dispose quant à elle d'outils permettant de la mettre à jour. La bibliographie est saisie de manière externe dans un fichier XML unique et est ensuite déposée sur la plate-forme. Au moment du dépôt, il est possible de vérifier le contenu des fiches bibliographiques et de générer une version téléchargeable. Un archivage des versions précédentes permet de revenir à un état antérieur. La gestion de la bibliographie s'effectue donc selon les mêmes principes que la rédaction des articles.

4 Conclusion

L'informatique est au coeur du projet DÉRom. La structuration des articles est formalisée par un balisage XML, qui améliore la pérennité des données, évite un fastidieux processus de rétroconversion des articles et, surtout, garantit l'homogénéité de la rédaction en contraignant la saisie des données. La rédaction via ce langage de balisage permet dès lors à de nombreux chercheurs de collaborer à ce projet depuis plusieurs universités et centres de recherche, tout en garantissant la cohérence du discours lexicographique. Dès leur approbation finale par les directeurs du projet, les articles sont publiés en ligne et rendus ainsi disponibles, sans délai, pour la communauté scientifique comme pour le

public averti. Cette publication en ligne permet en outre la mise à jour constante des articles après leur publication : le lecteur du DÉRom est ainsi assuré de consulter l'état le plus récent des recherches en étymologie romane. Enfin, la consultation en ligne des articles est facilitée par diverses fonctionnalités de recherche, offertes par la plate-forme ISIS, qui héberge, à l'ATILF, les projets lexicographiques de l'équipe « Linguistique historique française et romane ». L'intégration du DÉRom dans cet environnement nancéien en est encore à ses débuts ; gageons qu'elle permettra, à terme, sa mise en réseau avec d'autres ressources lexicographiques, d'abord internes à l'ATILF (nous pensons notamment au FEW, en cours d'informatisation : https://apps.atilf.fr/lecteurFEW), mais aussi externes, au premier chef desquelles le *Lessico Etimologico Italiano* (LEI, également en cours d'informatisation : http://woerterbuchnetz.de/LEI). Le moment venu, nous espérons que le DÉRom sera préparé à intégrer les propositions qui seront formulées au sein de l'action COST *European Network of e-Lexicography (ENeL)* (voir http://www.elexicography.eu), rejoignant ainsi le chantier européen en cours autour de ce qu'il est convenu d'appeler, depuis peu, la « lexicographie électronique » (Granger/Paquot 2012).

5 Bibliographie

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

Buchi, Éva/Gouvert, Xavier/Greub, Yan, *Data structuring in the DÉRom (*Dictionnaire Étymologique Roman*)*, in : Bettina Bock/Maria Kozianka (edd.), *Whilom Worlds of Words – Proceedings of the 6th International Conference on Historical Lexicography and Lexicology (Jena, 25–27 July 2012)*, Hambourg, Kovač, 2014, 125–134.

Dendien, Jacques/Pierrel, Jean-Marie, *Le* Trésor de la langue française informatisé. *Un exemple d'informatisation d'un dictionnaire de langue de référence*, TAL 43/2 (2003), 11–37.

DÉRom = Buchi, Éva/Schweickard, Wolfgang (dir.), *Dictionnaire Étymologique Roman (DÉRom)*, Nancy, ATILF, <http://www.atilf.fr/DERom>, 2008–.

DMF2012 = Martin, Robert/Bazin-Tacchella, Sylvie (dir.), *Dictionnaire du Moyen Français (DMF2012)*, Nancy, ATILF, <http://www.atilf.fr/dmf>, 2012.

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

Granger, Sylviane/Paquot, Magali, *Electronic Lexicography*, Oxford, Oxford University Press, 2012.

LEI = Pfister, Max/Schweickard, Wolfgang (dir.), *Lessico Etimologico Italiano*, Wiesbaden, Reichert, 1979–.

Martin, Robert, *Perspectives en lexicographie informatisée. L'expérience du DMF (*Dictionnaire du Moyen Français*)*, *Mémoires de la Société de linguistique de Paris* 7 (1999), 51–71.

Martin, Robert/Gerner, Hiltrud/Souvay, Gilles, *Présentation de la seconde version du DMF*, in : Maria Iliescu/Heidi Siller-Runggaldier/Paul Danler (edd.), *Actes du XXV^e Congrès International de Linguistique et de Philologie Romanes (Innsbruck 2007)*, Berlin/New York, De Gruyter, 2010, vol. 6, 213–220.

Renders, Pascale, *Modélisation d'un discours étymologique. Prolégomènes à l'informatisation du* Französisches Etymologisches Wörterbuch, Strasbourg, Société de linguistique romane/ÉLiPhi, à paraître.

REW$_3$ = Meyer-Lübke, Wilhelm, *Romanisches Etymologisches Wörterbuch*, Heidelberg, Winter, ^{3}1930–1935 [^{1}1911–1920].

TLF = Imbs, Paul/Quemada, Bernard, *Trésor de la langue française. Dictionnaire de la langue du XIX^e et du XX^e siècle (1789–1960)*, 16 vol., Paris, Éditions du CNRS/Gallimard, 1971–1994.

TLF-Étym = Steinfeld, Nadine (dir.), *Trésor de la langue française étymologique* , Nancy, ATILF, <http://www.atilf.fr/tlf-etym>, 2005–.