Grec ancien, latin médiéval, balisage comparé de deux dictionnaires, vers des ressources linguistiques

Frédéric Glorieux*— Sabine Thuillier**

** École nationale des chartes*
19, rue de la Sorbonne
75 005 Paris
frederic.glorieux@enc.sorbonne.fr

*** Projet Diccionario Griego-Español*
Instituto de Lenguas y Culturas del Mediterráneo y Oriente Próximo, CSIC
Albasanz, 26-28
28037 Madrid
sabine.thuillier@gmail.com

RÉSUMÉ. À partir de la présentation croisée du schéma d'encodage en XML/TEI de deux dictionnaires de langue ancienne, le Glossarium de Du Cange et le Diccionario Griego-Español dirigé par F.R. Adrados, nous observerons comment un même langage formel, la TEI, permet de tirer un parti renouvelé de ces ouvrages de référence quant à l'indexation du lexique du grec ancien et du latin médiéval.

ABSTRACT. Starting with a compared analysis of the XML/TEI tagging pattern of two ancient languages dictionaries - Du Cange's Glossarium, and the Diccionario Griego-Español, under the direction of Prof. F. R. Adrados - we observe how a single formal language, TEI, enables us to take new advantages of these dictionaries, in terms of lexical and usage indexation of Ancient Greek and Medieval Latin.

MOTS-CLÉS : lexicographie, TEI, grec ancien, latin médiéval.

KEYWORDS: lexicography, TEI, ancient greek, medieval latin.

Cet article présente comment l'informatisation d'un dictionnaire est une occasion de mieux le connaître. Il met aussi en parallèle deux projets de mise en ligne d'ouvrages imprimés : un glossaire **ancien** de latin médiéval, le *Glossarium mediae et infimae latinitatis* de Du Cange *et al.* (1re éd. 1678 ; dernière éd. 1875-1887)[1], et un dictionnaire **moderne** de grec ancien, le *Diccionario Griego-Español* (1980-)[2]. Ces références lexicographiques décrivent toutes deux une langue ancienne, comportent de nombreuses citations et s'étendent sur plusieurs volumes. Mais leur rapprochement est ici motivé par l'exploitation partagée du même standard d'encodage (les recommandations de la Text Encoding Initiative). À partir de divers échanges portant sur nos programmes d'informatisation respectifs, nous avons, en effet, été rapidement amenés à prendre une mesure significative des écarts et continuités entre deux lexicographies et deux traditions philologiques. Le croisement présenté dans les pages qui suivent s'opère donc à partir d'une description des formats (des ouvrages et de leur balisage informatique), et c'est le partage d'un même vocabulaire formel et la découverte de repères communs qui sous-tendent la présentation des deux dictionnaires.

Après avoir introduit le *Glossarium* et le *DGE* et présenté brièvement le format (XML-TEI) utilisé pour leur informatisation, nous analyserons ensuite les ouvrages en fonction de certains éléments de leur schéma d'encodage, en nous arrêtant sur trois ensembles qui correspondent à trois composantes essentielles du texte lexicographique : la lemmatisation, les citations, les significations. À chaque étape de cette description, nous tâcherons de mettre en perspective les richesses nouvellement exploitées de la lexicographie traditionnelle avec l'avancée du traitement automatique des langues anciennes.

[1] http://ducange.enc.sorbonne.fr/. Désormais *Du Cange* ou *Glossariun*.

[2] http://www.filol.csic.es/dge/index.htm . Désormais *DGE*.

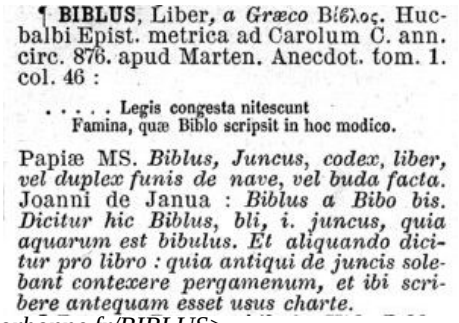
Table des matières

1. Présentation des ouvrages et de leur informatisation

1.1. Le *Glossarium mediae et infimae latinitatis*

Le *Du Cange* est un dictionnaire ancien, entrepris au XVII[e] siècle, et pourtant toujours consulté aujourd'hui, notamment par les médiévistes. D'autres outils sont parus depuis, comme le *Niermeyer*[3], mais la consultation de ce glossaire ancien reste encore un détour nécessaire pour la compréhension d'un texte médiéval.

Figure 1. *« BIBLUS » Glossarium mediae et infimae latinitatis, éd. augm., Niort : L. Favre, 1883 1887, t. II, col. 345a.*



¶ **BIBLUS**, Liber, *a Græco* Βίβλος. Hucbalbi Epist. metrica ad Carolum C. ann. circ. 876. apud Marten. Anecdot. tom. 1. col. 46 :

. . . . . Legis congesta nitescunt
Famina, quæ Biblo scripsit in hoc modico.

Papiæ MS. *Biblus, Juncus, codex, liber, vel duplex funis de nave, vel buda facta.* Joanni de Janua : *Biblus a Bibo bis. Dicitur hic Biblus, bli, i. juncus, quia aquarum est bibulus. Et aliquando dicitur pro libro : quia antiqui de juncis solebant contexere pergamenum, et ibi scribere antequam esset usus charte.*

<http://ducange.enc.sorbonne.fr/BIBLUS>

L'édition en ligne du *Glossarium* a été initiée par l'École nationale des chartes en 2007. Depuis 2009, l'Agence nationale de la recherche (ANR) finance le projet OMNIA, associant l'École des chartes, l'UMR 5594 ARTeHIS (université de Bourgogne - CNRS), et l'équipe de lexicographie latine de l'Institut de recherche et d'histoire des textes (IRHT). Le *Glossarium* numérisé sera complété par d'autres dictionnaires médiolatins en cours de rédaction à travers l'Europe, sous l'égide de l'Union académique internationale. Le *Novum Glossarium Medii Latinitatis* (*NGML*)[4], est le dictionnaire généraliste du latin médiéval (800-1200) rédigé par l'équipe de lexicographie de l'IRHT. Sur les fascicules publiés (L à Pl), deux sont en ligne[5] (de *phacoides* à *plaka*). Il s'agit à long terme de réunir sous un même portail tous les outils électroniques d'une lexicographie sur corpus : consultation des dictionnaires existants, rédaction collaborative, ainsi que des outils libres d'exploration de corpus étiquetés, dans l'esprit d'un Perseus[6] pour le latin médiéval.

[3] Niermeyer, *Mediae latinitatis lexicon minus, Lexique latin médiéval français-anglais ; a Medieval Latin French-English Dictionnary*, Leyde, E. J. Brill, 1954-…

[4] http://www.irht.cnrs.fr/recherche/lexico.htm

[5] http://omnia.enc.sorbonne.fr/

[6] Université de Tufts, Perseus *digital library* http://www.Perseus.tufts.edu/.

1.2. Le *Diccionario Griego-Español*

Le *DGE* est un dictionnaire bilingue monodirectionnel (grec-espagnol) initié et dirigé depuis 1962 par le linguiste Francisco Rodríguez Adrados, au sein du Consejo Superior de Investigaciones Científicas (Madrid). Le premier volume a été publié en 1980 et le volume VII est paru en 2009, la dernière entrée en est ἔξαυος. Reprenant le flambeau de la longue tradition de la lexicographie grecque, l'ambition première du *DGE* est de contribuer au renouvellement de cette dernière, en proposant aux lecteurs hispanophones un ouvrage de référence mettant à jour et enrichissant son prédécesseur immédiat et modèle, le *Greek-English Lexicon*, plus connu sous le nom de ses éditeurs, Liddell-Scott-Jones (LSJ)[7]. Face au développement exceptionnel des ressources philologiques – papier et électroniques – le projet espagnol a en effet l'ambition de traiter de la manière la plus exhaustive possible le lexique grec et d'étendre son corpus au plus grand nombre de textes (textes littéraires, commentaires, inscriptions, papyrus sublittéraires et documentaires, sur une période allant du mycénien au VI[e] siècle après J.C, limite étendue au X[e] siècle pour certaines parties du corpus).

βύβλος, *-ου, ἡ* **βίβλ-** Hermipp.63.13, D.18. 259, 19.199, Thphr.*HP* 4.8.4, **βύβλον** *AP* 9.98 (Stat.Flacc.) [*-ŭ-*] **I** bot. **1** *papiro, Cyperus papyrus* L., utilizado como alimento *βύβλου καρπός* A.*Supp.*761, cf. Hdt.2.92, Str.17.1.15, Phryn. 270; esp. el *tallo* apreciado por los muchos usos de su fibra *ἐκ τῆς βίβλου ἱστία τε πλέκουσι* Thphr.l.c., cf. Hdt.2.96, *τιμὴ βίβλου μυριάδων δύο el precio de 20.000 papiros*, *PTeb*.308.7 (II d.C.). **2** *στεφανωτρὶς β. papiro coronario* variedad de papiro utilizado para hacer coronas Theopomp. Hist.106, 107.

Figure 2. *DGE, extrait de l'article* βύβλος.

Le projet d'informatisation actuellement en cours au sein du *DGE* consiste à mettre en valeur un texte fortement structuré et régulier, à l'aide d'un encodage en XML TEI, permettant des fonctionnalités plus avancées de consultation et de recherche en ligne. En guise de première expérience visant à valider le schéma et le balisage appliqués aux fichiers du *DGE*, le projet espagnol a procédé à l'informatisation du *Léxico de magia y religión en los papiros mágicos griegos* (*LMPG*), publication annexe du *DGE*, parue en 2001[8]. Cet ouvrage, rédigé par Luis Muñoz Delgado, est de moindre ampleur (2 627 entrées, 400 pages) et analyse sémantiquement un corpus plus homogène et plus restreint (les papyrus magiques grecs du IIe siècle av. J.-C. au Ve siècle après J.C. recueillis notamment dans les *Papyri Graecae Magicae* et le *Supplementum Magicum*[9]). Il présente en outre l'intérêt de suivre strictement la structure et la mise en forme du *DGE*. La mise en ligne de cet ouvrage a été finalisée cette année grâce à l'importante contribution de l'équipe informatique de l'École des chartes responsable de l'informatisation du *Du Cange*. Nous pouvons d'ailleurs remarquer ici que le travail commun qui a pu être engagé à cette occasion entre nos deux équipes a été largement facilité par le partage d'un même « lexique » : le vocabulaire des balises TEI décrivant la structure d'un dictionnaire.

1.3. XML et TEI

Il convient sans doute d'expliquer en quelques mots, pour terminer notre présentation, ce que sont XML et la TEI, et la raison pour laquelle ils constituent un format adapté à l'informatisation de dictionnaires.

XML (*eXtensible Markup Language*)[10] est une syntaxe de balisage permettant de segmenter un texte et d'accrocher des étiquettes (éléments descriptifs entre chevrons) à ces segments. Ceux-ci peuvent s'imbriquer et constituer des « arbres ». L'ordre du flux textuel est donc conservé, les balises permettent d'encoder tous les événements de la typographie.[11] De plus, XML est un langage dit *eXtensible* car il ne précise pas le vocabulaire à inscrire entre les crochets

[7] Depuis sa numérisation SGML en 1994 par le projet Perseus, le LSJ, dans l'édition de 1940, est en ligne gratuitement. Perseus offre dorénavant ses ressources sous une licence libre, et d'autres sites proposent le LSJ en ligne, notamment l'université de Chicago avec le logiciel Philologic de Mark Olsen, http://www.lib.uchicago.edu/efts/PERSEUS/Reference/lsj.html, ainsi qu'une version très simple et fort pratique : http://philolog.us/.

[8] http://dge.cchs.csic.es/lmpg/

[9] Respectivement édités par K. Preisendanz et R. Daniel - F. Maltomini.

[10] http://www.w3.org/XML/

[11] Signalons à ce propos que le *DGE* est rédigé avec le logiciel commercial *WordPerfect* (Corel) depuis les années 1990. Ce produit se distingue par son langage de macro, *PerfectScript*, permettant l'automatisation de nombreuses corrections typographiques. Il en résulte que la ponctuation « visible » est strictement vérifiée, ce qu'a confirmé la conversion en XML (également effectuée avec des macros *WordPerfect*). Un exemple de cette méthode de conversion : toutes les entrées ont été balisées avec un script formulant à peu près ceci : « repérage du début de paragraphe avec un retrait (une tabulation), puis, en gras et en alphabet grec, lemme principal, pouvant être précédé par : 1) un numéro arabe pour distinguer des homographes, 2) un obèle (†) pour un mot signalé comme corrompu, 3) par un astérisque pour un mot reconstitué du syllabaire mycénien ». Cette expérience permet de rappeler que la production d'XML ne nécessite pas de compétences informatiques rares, même un logiciel bureautique peut générer un document structuré, à la condition d'avoir une définition intellectuelle précise, un schéma.

obliques (les balises). Un document peut employer des *éléments* : `<dictionnaire>`, `<article>`, `<définition>`, `<exemple>` ; tout aussi bien que `<html>`, `<div>`, `<h1>`, `<p>` ; ou bien encore `<entry>`, `<form>`, `<sense>`. La « ponctuation » XML permet ainsi à chaque projet de définir son vocabulaire de balises, et la grammaire qui règle leurs successions et leurs enchâssements : un schéma.

Pour le *Du Cange* comme le *DGE* (et bien entendu le *LMPG*), nous avons choisi d'écrire manuellement un schéma spécifique, dans une grammaire exactement contrôlée, mais nous avons décidé de suivre un vocabulaire standardisé, celui recommandé par la *Text Encoding Initiative* (TEI) [12]. Le choix d'un vocabulaire de balises pose en effet des questions terminologiques : quelle langue choisir pour un nom ? Faut-il dire *article*, *entrée*, ou *articulus* ? Fallait-il des éléments en latin pour le *Du Cange*, en grec ou en espagnol pour le *DGE* ? Une convention comme la TEI ne les évacue pas totalement, les noms et les notions implicites seraient discutables, mais une fois inscrits dans le code, ces mots univoques aident à quantifier, à comparer et à échanger ses sources et méthodes d'un projet à l'autre. De plus, la TEI est un standard acamédique qui se consacre à l'édition et au partage des textes de sciences humaines et sociales, et dédie un chapitre de ses *Guidelines* (*Recommandations*) au balisage des dictionnaires[13]. Une fois encore, ce sont ces références communes qui ont ici motivé la rencontre de deux dictionnaires, de deux lexicographies, de deux langues anciennes et l'on peut comprendre alors dans quelle mesure les élément TEI deviennent, comme déjà évoqué au début de ce propos, des repères partagés entre le *Du Cange* et le *DGE*, et appuieront dorénavant la mise en écho décrivant le contenu des deux ouvrages et leurs exploitations envisagées pour des traitements automatisés du grec ancien et du latin médiéval.[14]

2. Macrostructure : le dictionnaire comme base d'unités lexicales, <entry>

Un dictionnaire repose sur un atomisme qui divise le lexique en lemmes. Ce modèle logique convient bien à l'informatique. L'agencement alphabétique et lemmatisé du lexique d'une langue ancienne, préalable à toute entreprise lexicographique, exige toutefois une fine expertise philologique afin de dénouer l'écheveau de la variation graphique sur les siècles et les territoires. Comment baliser et corriger la nomenclature pour prévoir d'alimenter les bases lexicales nécessaires à des lemmatiseurs, en déjouant la variation graphique des langues anciennes ?

2.1. Nomenclature : hiérarchiser les lemmes et les variantes, <form>, <orth>, <ref>

Les lecteurs du *Glossarium* savent que l'ouvrage est peu systématique, l'informatisation a mesuré jusqu'à quel point. Des tests automatisés signalent que 1 500 vedettes sur 90 000 ne suivent pas l'ordre alphabétique. Certains cas s'expliquent par des équivalences de graphèmes, tels 'U' et 'V', 'I' et 'J', ou 'Æ' et 'E'. La succession INCLAVATURA, INCLAUDARE, INCLAUDERE, INCLAVELARE respecte donc l'ordre de l'alphabet latin. La confusion résulte aussi de l'histoire éditoriale de l'ouvrage. Au fil des siècles, les lexicographes ont inséré leur propos dans de nouveaux articles, ou des sous-articles, dans un souci plus visuel que philologique. Un mot en capitales grasses identifie normalement une vedette, tandis qu'une sous-vedette devrait se présenter en petites capitales. Une segmentation automatique selon la typographie ne distingue par exemple qu'un seul article « 1. BREVIS, BREVE ». Mais pourquoi la vedette est-elle numérotée alors que l'article est unique ? Un examen de la page imprimée découvre treize autres BREVIS ou BREVE, précisément numérotés, mais en petites capitales. Des scripts ont permis de repérer ce type d'irrégularités, produisant de nombreux faux homographes. Par ailleurs, les graphies médiévales ne sont pas fixées, la nomenclature répertorie de nombreuses variantes. Considérez par exemple ce parcours : « AYSSARTARE, pro Essartare », *essartare* n'est pas à la nomenclature, mais derrière *essartum*, « ESSARTUM, Essartare. Vide Exartus. », « EXARTUS, Exartum, Exartes, Essartum, Assartum, Sartus, Sartum », « SARTUM, vel Sartus ». L'article AYSSARTARE comporte deux paragraphes, EXARTUS en regroupe vingt, SARTUM ajoute encore deux citations. Le système d'adressage du *Glossarium* ne permet donc pas, en l'état, d'établir une hiérarchie rigoureuse entre variante et lemme, compliquant la consultation, et surtout l'extraction de ressources linguistiques. L'équipe OMNIA, avec les lexicographes du *NGML*, procède à l'établisse-

[12] http://www.tei-c.org/

[13] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html. Voir aussi Ide, N - Véronis J. (1996).

[14] Pour une description détaillée de ces éléments, nous renvoyons les lecteurs à la documentation publiée sur le site TEI, http://www.-tei-c.org/release/doc/tei-p5-doc/fr/html/REF-ELEMENTS.html.

ment d'une nomenclature de référence pour un lemmatiseur. Le réseau des entrées et des renvois du vieux *Glossarium* est soumis à la vérification philologique pour alimenter des listes de lemmes et de variantes. Cette nomenclature sert d'abord à la navigation entre les articles et à rassembler sous un seul lien les articles supposés porter sur un même mot.

Le grec ancien n'est certes pas épargné par ces phénomènes de variations et les flottements qui en résultent, mais les rédacteurs du *DGE* peuvent s'appuyer sur une longue tradition très documentée de l'étude des faits morphologiques, qui caractérise l'histoire de la linguistique grecque. Cela permet aux lexicographes d'établir la hiérarchie entre lemme et variante en les regroupant sous le même article. Concession au lecteur et au papier, quand une variante est alphabétiquement éloignée de son lemme, elle est indiquée par un renvoi, tout en étant systématiquement rappelée dans l'article. Le *DGE* propose un système d'adressage plus facile et plus sûr à informatiser que celui du LSJ[15]. Sa nomenclature est en outre l'exemple d'une indexation particulièrement précise et complète du vocabulaire grec. Cette visée d'exhaustivité chère au projet espagnol ne dispense pas d'un effort critique, bien au contraire[16] : il ne s'agit pas de retenir toutes les formes découvertes sans un examen linguistique et philologique scrupuleux, et de ne mentionner que les variantes morphologiques non régulières. Le choix a été fait de ne séparer comme deux mots distincts que les formes étymologiquement et sémantiquement distinctes. Quand la différence entre deux formes, purement phonétique ou graphique, tient à des raisons d'ordre chronologique ou dialectal, les formes sont considérées comme appartenant à un même article (ou mot). Si une forme attique est attestée, elle constitue, à de très rares exceptions près, le lemme principal.

Le travail de regroupement morphologique ayant déjà été assuré par les lexicographes, le repérage informatique a donc été beaucoup plus aisé que pour le *Glossarium* : l'entrée s'ouvre sur un en-tête morphologique, `<form>`, contenant la vedette, `<orth type="lemma">`, suivie de ses éventuelles variantes graphiques, phonétiques, morphologiques et dialectales, et d'informations prosodiques. Chaque variante, `<orth type="variant">` ouvre un nouvel élément `<form>`, imbriqué dans l'adresse principale : `(form (orth, gramGrp, form*))`[17]. Les articles de renvoi, contenant les variantes alphabétiquement éloignées du lemme principal, sont encodés comme une référence croisée, `<xr>`, pointant vers l'entrée principale, `<ref>`.

```
<entry>
 <form>
  <orth type="lemma">ὐ</orth> …
  <form><orth type="variant">ἰ -</orth>…</form>
  <form><orth type="variant">ὐ</orth>…</form>
  <form type="pros">-ῠ-</form>
 </form> …
</entry>

<entry>
 <form>
  <orth> -</orth>
 </form>
 <xr>v. <ref>u -</ref></xr>
</entry>
```

La précision du balisage du *DGE*, grâce à l'indexation morphologiquement structurée des lemmes et de ses variantes, permet de disposer d'une nouvelle source fiable de lemmes pour corriger et enrichir les bases lexicales du grec ancien actuellement disponibles. Pensons à l'ample projet Perseus[18] qui a développé un programme d'analyse morphosyntaxique, *Morpheus*, contenant une base lexicale qui a d'ailleurs servi au projet de lemmatisation du *Thesaurus Linguae Graecae* en ligne[19]. Malgré les grandes avancées du programme, subsistent encore diverses difficultés, des erreurs,

---

[15] Sur cette question du regroupement morphologique, signalons que le LSJ allait en effet jusqu'à rassembler sous un même paragraphe des familles lexicales plus ou moins dérivationnelles – disposition d'ailleurs abandonnée par les versions électroniques, mais parfois un peu imprudemment : des lemmes mal accentués ont été recomposés par des automates avec trop peu de vérifications. Ces regroupements n'avaient pas de vraies assises linguistiques, la valeur scientifique de ces classifications était faible : il s'agissait surtout de gagner de la place sur le papier, en tronquant le début de certains lemmes, et non d'ordonner systématiquement et hiérarchiquement les familles lexicales, d'où l'impossibilité d'en dégager des données exploitables pour la formation des mots en grec.

[16] Cf. H. Rodríguez Somolinos, « El *DGE* y la epigrafía griega: el problema de las palabras fantasma (ejemplificación y tipología) », *Miscelánea léxica en memoria de Conchita Serrano*, *Manuales y Anejos de Emerita* 41, Madrid, CSIC, 1999, 187-198, et J. Rodríguez Somolinos - J.A. Berenguer, « Lexicographie grecque et papyrologie. Le *Diccionario Griego-Español* », *Akten des 21. Internationalen Papyrologenkongresses- Berlin, 13.-19.9.1995*, *Archiv für Papyrusforschung Beiheft* 3, Stuttgart - Leipzig, Teubner, 1997, pp.858-866. Ces deux articles sont repris dans Adrados, F.R. - Rodríguez Somolinos, J. (eds) (2005).

[17] Cette pseudo syntaxe s'inspire du langage DTD et de Relax-NG Compact.

[18] http://www.Perseus.tufts.edu/hopper/

[19] http://www.tlg.uci.edu/. Le *TLG* a poursuivi le projet de lemmatisation, et annonce à présent une reconnaissance automatique de 93.2% de son corpus. Mentionnons aussi le Dictionnaire Automatique Grec (DAG) développé au sein du Projet de Recherche en Lexicologie Grecque (U.C.L.). Cette base lexicale, accessible aux chercheurs et élaborée à partir de concordances lemmatisées du corpus patristique et byzantin compte 306.814 formes attachées à 62.229 lemmes.

des mots élidés, autant d'occurrences non ou mal résolues. La prise en compte des formes présentes dans le *DGE* est un moyen de contrôler des mauvaises interprétations morphologiques effectuées par l'automate[20].

En résumé, si la nomenclature strictement établie du *DGE* permet de corriger les bases lexicales du grec ancien, la communauté médiolatine n'est pas aussi avancée ; le *Du Cange* fournit une matière, mais pas de structure.

2.2. Grammaire : normalisation des données, <gram>, <pos>

La base lexicale nécessaire à un lemmatiseur commence par une nomenclature de lemmes (et de variantes), s'y accrochent ensuite une étiquette morphosyntaxique (nom, verbe, adjectif…), afin de relier le lemme à ses formes fléchies. Le dictionnaire épargne beaucoup de saisie lorsqu'il fournit des indications grammaticales normalisées[21] de nature et de paradigme flexionnel (déclinaison ou conjugaison).

En plus des données de variations, l'en-tête morphologique des articles du *DGE* indique, comme il est d'usage en lexicographie de langue flexionnelle, les motifs identifiant le paradigme de déclinaison. Ces données sont balisées par des éléments `<gram>` encadrés par un `<gramGrp>` et l'étape actuelle d'informatisation a permis d'ajouter dans l'entête une balise `<pos>`, étiquette mopho-syntaxique déduite :

```
<form>
 <orth>διασφράξ</orth>
 <gramGrp>
  <pos value="SF"/>
  <gram type="genitive">-άγος</gram>
  <gram type="determiner">ἡ</gram>
 </gramGrp>
</form>
```

L'objectif premier de cet étiquetage grammatical des lemmes du *DGE* n'est pas le traitement de la langue – il existe déjà de grands projets d'analyseurs – mais l'exploration du dictionnaire. Une application exploitera cette information pour autoriser des requêtes croisées telles que celle-ci : « quels sont les entrées décrivant un adjectif et comportant une citation d'Homère ? » (ce que le LSJ électronique ne permet pas).

En revanche, le *Du Cange* ne contient aucune information grammaticale. Il ne répertorie pas non plus de mots grammaticaux, et très peu de verbes. Le glossaire compile surtout les substantifs et les adjectifs révélant des réalités médiévales inconnues du latin classique. Cette néologie a des graphies fort diverses, mais, heureusement, pas autant de variété flexionnelle que le latin classique. Les déclinaisons se déduisent généralement de la désinence du lemme ou de la variante. Les conjugaisons et les déclinaisons les plus défectives proviennent généralement du latin classique, déjà décrit et répertorié (base lexicale Perseus).

Le lien entre formes et lemmes étiquetés peut s'automatiser de deux manières. Soit un raciniseur (*stemmer*) part de formes attestées pour rechercher le ou les lemmes les plus probables, soit un fléchisseur part des lemmes étiquetés du lexique et produit toutes les formes déclinées et conjuguées possibles. Dans la pratique, ces principes généraux s'associent sous une supervision humaine, mais ils permettent d'éclairer une distinction formelle. Pour le grec ancien (ou le latin classique) le corpus est clos, presque entièrement numérisé[22], il est inutile de générer des formes qui ne se rencontreront plus jamais. Pour le latin médiéval, le corpus conservé est plus important, moins numérisé, ni même édité, la base lexicale est ouverte, d'autant plus que la créativité lexicale et morphologique des médiévaux est imprévisible. Des textes théologiques pourrait ainsi demander un analyseur dérivationnel, afin de réduire les concepts forgés avec force préfixes et suffixes.

[20] Deux exemples : 1. ἐπέμφθ' : forme élidée de l'aoriste passif ἐπέμφθη (πέμπω). L'index du site du *TLG* la mentionne, mais l'analyseur ne renvoie pas à πέμπω, mais interprète ἐπέμφθ′ comme l'aoriste actif de deux lemmes fantômes *ἐπεμφθάνω et *ἐπεμπέτομαι. 2. βούκολον, -ου, τό : substantif neutre, recensé dans le *DGE* et absent du LSJ, attesté par deux fois dans les *Oxyrhynchus Papyri*, présents dans la librairie de Perseus, mais que *Morpheus* ne connaît pas.

[21] Soit « Bellus, a, um », un motif comme *–us, -a, -um* désigne sans équivoque un adjectif et permet de programmer un fléchisseur qui génère toutes les formes de la déclinaison. Si une grande partie du lexique est régulière, il existe de nombreux exceptions, ainsi que des erreurs, ou simplement des variations humaines : « edo, edis ou es, edere ou esse, edi, esum ».

[22] Notamment grâce aux importantes banques d'inscriptions et de papyrus en ligne, et celle du *TLG*, dont la dernière version (en ligne) contient 3 800 auteurs, 12 000 œuvres et environ 99 millions de mots. La collection grecque de Perseus n'en contient « que » 8 millions à peu près, mais elle est librement accessible.

3. Citations : le dictionnaire comme réseau dans le corpus des textes, <cit>

En plus de mettre en ordre morphologiquement les unités lexicales, décrire le lexique d'une langue ancienne revient, pour le lexicographe comme pour le lexicologue, à examiner et à choisir des acceptions remarquables dans un corpus textuel. Le dictionnaire est ainsi non seulement un réseau avec ses liens internes (les renvois de formes dans la nomenclature), mais aussi ses liens externes (les citations référencées). Le *Du Cange* et le *DGE* citent beaucoup[23] ; chaque article, s'il ne s'agit pas d'une référence croisée, comporte au moins une citation. L'article d'un dictionnaire élaboré sur corpus pourrait se concevoir comme une *relation*, au sens formel de l'algèbre de Codd (SQL), entre le lexique et la bibliographie. Mais quels traitements sont nécessaires pour exploiter ces citations référencées ? Nous allons voir que les exploitations envisageables divergent fortement d'un dictionnaire à l'autre.

3.1. De la concordance à la lexicographie de corpus

Que se soit pour Homère ou la Vulgate, les premiers glossaires collectaient les notes marginales pour éclairer un mot du texte manuscrit. L'idéal de cette description lexicale a été atteint par Hugues de Saint Cher qui établit en 1230 la concordance complète de la Bible. Tous les mots du texte étaient répertoriés, avec les références de chaque occurrence. La complétude renversait la perspective : ce n'était plus le texte qui gouvernait les mots, mais l'ordre alphabétique qui organisait les citations, c'est en quelque sorte le premier dictionnaire de corpus.

La taille et l'éparpillement du corpus médiéval ne permet pas cette complétude. Un dictionnaire n'y est pas encore un recensement, mais plutôt, un sondage. Du Cange commença son glossaire en notant les « barbarismes » qu'il rencontrait dans les documents d'archives, les éditions imprimées, sur des fiches rangées en ordre alphabétique. La méthode est restée la même jusqu'à l'informatique. Littré ne procédait pas autrement, embauchant des étudiants pour lire les Classiques en y répertoriant des emplois remarquables. Les locaux du *NGML* à l'Institut de France sont encore tapissés du fichier commencé en 1923 pour le dépouillement des sources. Si le dictionnaire peut promettre de rendre compte de toutes les fiches, l'aléa se déplace sur la constitution du fichier, et les textes accessibles à une époque. Le *Glossarium* semble composé comme une suite d'extraits glosés selon l'ordre alphabétique, sans aucune considération de fréquence des occurrences, avec la passion du collectionneur pour principe. La matière conservée n'a pas été mutilée par une doctrine, mais elle n'est pas pondérée, ni même définie, sinon par l'esprit de finesse de l'âge classique.

Le *DGE*, lui, obéit aux principes de la lexicographie de corpus[24]. Les rédacteurs poursuivent l'attentif examen critique de leurs prédécesseurs et disposent, en plus des articles des dictionnaires précédents, d'un matériel lexicographique très important, issu pour l'essentiel du dépouillement, opéré au sein du laboratoire, des nouvelles éditions critiques, de leurs index, des nouvelles collections d'inscriptions et de papyrus. Chaque nuance de sens ou acception est toujours accompagnée d'au moins une citation, `<cit>`, incluant le plus souvent un extrait de texte cité, `<quote>`, et en tous cas toujours une référence bibliographique, `<bibl>`. Selon la pertinence et la valeur exemplaire du contexte immédiat de l'occurrence en mention, le *DGE*, suivant là encore la méthode de son prédécesseur, le LSJ, peut ne faire que renvoyer au texte, sans le citer ni le traduire. En plus de la référence, les extraits sont parfois suivis d'une traduction en espagnol. Ceci peut s'exprimer de manière formelle ainsi : `cit (quote?, trans?, bibl+)`. Les critères de sélection sont multiples (chronologie, niveau de langue, genre littéraire, fréquence, etc.), et en présenter ici le détail nous porterait loin de notre propos. Les choix effectués reposent sur l'intention de témoigner de la variété des faits linguistiques attestés sur une période de 1500 ans, en se gardant d'en dégager une norme.

3.2. Extraits : les citations du dictionnaire comme corpus de haute variation, <quote>

« En tant que clé d'accès au texte, la typographie est bien entendu essentielle. (…) Dans les dictionnaires (…), il n'y a pas d'application, au sens mathématique, entre l'ensemble des champs informationnels et l'ensemble des styles disponibles, un même champ informationnel pouvant revêtir plusieurs styles. »[25] Cet énoncé de mathématique ensembliste (fonctions, applications, bijections...) peut se reformuler dans une logique plus classique, afin d'expliquer un problème critique du *Glossarium* : toutes les citations ne sont pas en italique, certains extraits en vers paraissent en caractères romains plus petits ; tout ce qui est italique n'est pas citation, c'est aussi la distinction typographique des mots étrangers, des mots en mention, des renvois, et bien d'autres nuances conventionnelles. Il en résulte que, dans l'état actuel de la source balisée du *Du Cange*, 160 000 citations supposées ont été balisées `<quote>`. Un filtre automatique a interprété tous les segments italiques précédés de deux points ':' comme des citations introduites par des références. La règle semble de bon rendement, mais 8 000 `<quote>` douteux (5%) ont déjà été détectés et demandent à être relus. Une

[23] Il est prévu que le nombre de citations du *DGE* sera multiplié par trois par rapport au LSJ. Pour le matériel papyrologique et épigraphique, l'enrichissement bibliographique est encore plus considérable.

[24] Les listes des auteurs et œuvres cités par le *DGE* ainsi que les éditions de références utilisées sont consultables sur le site du projet : http://www.filol.csic.es/dge/lst/3lst-int.htm.

[25] Wionet Chantal, Tutin Agnès, *Informatisation du* Dictionnaire universel *de Furetière revu par Basnage (1702) : premier bilan*, Paris, Champion, 2001.

citation du *Glossarium* peut en effet avoir trois niveaux de hiérarchie. Soit par exemple dans l'article DESVIATORIUM cette citation en ancien français : « *Jehan Pigasse avoit fait aucunes destrousses et excluses* (rectius infra : *excluses* et *Destournées*) *dedans le pré d'iceulx Crosmanas, pour oster l'eaue de leur pré.* ». La citation est en italique, elle contient une note du lexicographe en caractères romains et entre parenthèses, cette note indique deux mots italiques en mention, *excluses* et *Destournées*, qui ne font pas partie de la citation. On comprendra la difficulté de programmer des automates qui non seulement prévoient tous les cas, mais aussi corrigent les erreurs humaines inévitables.

Les extraits cités par le *DGE*, grâce à la différence d'alphabet, se repèrent et se délimitent bien plus facilement : ils correspondent aux segments de texte grec qui se trouvent après l'en-tête morphologique et avant la section étymologique. Étant donnée l'abondance, déjà évoquée plus haut, des sources textuelles numérisées, aucune exploitation plus avancée de ces `<quote>` n'est engagée. En effet, dans la problématique de l'exploration automatisée de corpus, le grec ancien est dans la situation particulièrement favorable où presque toute hypothèse linguistique peut être vérifiée sur la totalité des occurrences conservées de la langue ; le *DGE* est donc avant tout une organisation sélective de la masse de ces témoignages, et l'enjeu est avant tout de relier les références de ces citations aux corpus numérisés disponibles.

3.3. Références : des conventions bibliographiques au lien Internet, <bibl>

Dans l'idéal, un dictionnaire électronique sur corpus devrait pouvoir être lié aux textes qu'il cite. Le *LSJ* en ligne sur le site Perseus est un modèle : 422 000 citations ont été balisées et, pour les textes faisant partie de la librairie électronique du projet, ces références font ainsi office de liens pointant vers le paragraphe exact de l'œuvre citée. L'objectif technique consiste à convertir des références bibliographiques en *URI*[26] et le degré d'automatisation de cette tâche dépend pour beaucoup du degré de la normalisation de ces références.

Ainsi, relier le *Du Cange* à son corpus reviendrait à pouvoir inscrire une balise `<bibl>` et un attribut `@xlink:href`.

```
<cit>
 <bibl xlink:href="???">Charta Th. decani S. Vulfr.
 Abbavil. ann. 1218. ex primo Lib. nigr. ejusd. eccl.
 fol. 8. r°.</bibl> :
 <quote>Retenta sibi et hæredibus suis præpositura cum
 dominio et libertate et fructibus grangiæ per servicium,
 quod antea nobis reddere solebat, videlicet duellum, et
 Citationes, et alia servicia.</quote>
</cit>
```

Pour le *Glossarium*, une référence bibliographique, `<bibl>`, serait *a priori* la phrase qui précède une citation. Toutefois, la quantité de points et d'abréviations non référencées perturbe la plupart des segmenteurs. Le lexicographe n'a pas jugé nécessaire de suivre une convention de ponctuation pour isoler les références. L'attribut de lien, `@xlink:href`, est encore plus difficile à renseigner. « Charta Th. decani S. Vulfr. Abbavil. ann. 1218. ex primo Lib. nigr. ejusd. eccl. fol. 8. r°. » correspond à un manuscrit connu des médiévistes, Bibliothèque nationale de France, nouvelles acquisitions latines 1681, fol. 8, cartulaire dit « livre noir » de la collégiale d'Abbeville, identifiable dans la base *CartulR* de l'Institut de recherche et d'histoire des textes, à l'adresse `http://www.cn-telma.fr/cartulR/entite5363/`. Les médiévistes risquent d'être encore longtemps les seuls à savoir lire ces références.

Ces difficultés rappellent ici encore que la profondeur d'une informatisation est largement tributaire des siècles de normalisation qui précèdent. En philologie grecque, la normalisation des références et des abréviations bibliographiques est très ancienne et bien plus régulière que pour la plupart des autres langues, modernes et anciennes. Ainsi, par exemple, la résolution des références du corpus platonicien est possible parce qu'Henri Estienne a établi un modèle avec l'édition de référence de Platon (Genève, 1578), qui donne encore le patron de numérotation des citations de l'œuvre. La papyrologie a aussi établi très tôt ses propres canons.[27] Il est donc possible, pour l'encodage du *DGE*, de tirer parti des sous-éléments de `<bibl>` proposés par la TEI et de segmenter les références selon l'auteur, `<author>`, l'œuvre `<title>`, et la localisation du passage en mention, `<biblScope>` :

```
εἰρεσία, -ας, ἡ  ... I ... E.Hel.1453
<bibl type="related">
 <author>E.</author><title>Hel.</title><biblScope>1453</biblScope>
</bibl>
```

[26] *URI* : *Unique Resource Identifier*, les identifiants sur Internet permettant d'établir des liens.

[27] Les références papyrologiques, attentivement suivies par le *DGE*, sont régulièrement mises à jour par l'American Society of Papyrologists, http://scriptorium.lib.duke.edu/papyrus/texts/clist.html.

Le degré de détail qui est ainsi atteint dans le balisage des références ouvrira les recherches croisées à un auteur, une œuvre, etc.

Par ailleurs, le *DGE* a pris le parti de réviser systématiquement toutes les citations du LSJ et de les reprendre, s'il n'existe pas de meilleure lecture du texte cité. Est aussi adopté, autant que possible, le même système d'abréviations : le texte cité est parfois différent car vérifié sur des éditions plus récentes, mais la référence reste souvent la même. Le *DGE* en ligne pourra donc récupérer, pour les citations communes, les liens établis par Perseus entre le LSJ et les corpus[28]. Ces relations pourront aussi s'étendre à d'autres corpus de textes cités par le *DGE* et accessibles en ligne, tels que la Suda On Line[29], les grandes banques de textes papyrologiques du Duke Databank of Documentary Papyri et de la Heidelberger Gesamtverzeichnis[30], ou encore le corpus d'inscriptions du Packard Humanities Institute Greek Epigraphy Project[31]. On peut d'ailleurs attendre de véritables avancées dans ce champ de publications, grâce, entre autres, au projet EpiDoc qui vise à diffuser un standard d'encodage XML pour l'édition électronique du matériel épigraphique et papyrologique à partir de la TEI.[32]

En résumé, la qualité et l'ampleur du corpus antique référencé permet au *DGE* de proposer une organisation sélective de citations selon l'intention scientifique consciente du projet éditorial. En revanche, les extraits cités par le *Glossarium* ne sont, eux, pour l'instant représentatifs que de la variété du lexique[33].

4. Microstructure : formaliser l'arbre sémantique, <sense>

L'article d'un dictionnaire de langue ancienne pourrait-il n'être qu'une suite de citations sans gloses ni traductions ? Un concordancier avec des fonctionnalités avancées de tri et de statistiques sur les cooccurrents aurait l'avantage de ne pas masquer les extraits qui ne rentrent pas dans le plan du lexicographe. Les lexicographes du *NGML* ont envisagé l'hypothèse mais constatent : « [Les CD-Roms] embrassent de considérables corpus de données ; mais il s'agit, en quelque sorte, d'un matériau brut, malaisé à utiliser, en particulier lorsque l'on travaille sur des vocables de fréquence élevée ou très élevée. […] Les dictionnaires offrent au contraire des attestations limitées en nombre, mais un matériau déjà élaboré : pour une information rapide sur les valeurs sémantiques de tel ou tel vocable, ils sont et resteront certainement des instruments très commodes. »[34] À l'heure informatique, l'arbre des significations d'un dictionnaire conserve une utilité pour le lecteur humain, comme le plan d'un livre permet d'entrer plus profondément dans l'intelligence de l'auteur et son sujet. Est-ce que cet arbre, format computationnel avec ses algèbres et ses algorithmes, peut avoir des exploitations autres que visuelles, par exemple pour la désambigüisation automatique de mots très polysémique ? Probablement pas à court terme, car les critères utilisés pour distinguer un A d'un B dans le dictionnaire sont conduits par la communication entre humains, selon un bon sens que la machine ne partage pas. Cependant, d'autres champs d'information peuvent être exploités, comme les traductions et les marques d'usage, afin d'alimenter des bases lexicales avec des informations syntaxiques et sémantiques.

4.1. Plan : du glossaire au dictionnaire bilingue, <dictScrap>, <sense>

Décrire la composition d'un article du *Glossarium* commence par une énumération de négations afin de prévenir les déceptions du lecteur : un glossaire n'est pas un dictionnaire, rien n'est systématique. On n'y trouve pas de définitions à proprement parler, mais des gloses hétérogènes, souvent courtes, parfois de vraies dissertations, toujours écrites en latin. Les mots français que l'on y reconnaît sont rarement des traductions, et plus souvent une origine étymologique. En effet, nombre de mots du latin médiéval proviennent de l'ancien français, de l'italien, du germanique, des langues vernaculaires en général. « 1. CARCER SUB TERRAM [...] Nostris *Cul-de-bassefosse*. », « HEUÇA, Heusa, a veteri Gallico *Heuce* vel *Heus* et *Heuse* [...] » (une cheville de métal), « PIZZA, Placenta, ex Italico *Pinza* ». Pour l'informatisation, le corps d'un article, `<entry>`, n'a pas plus de structure régulière et segmentable qu'une suite de paragraphes. La richesse réelle mais hétérogène des articles du *Glossarium* ne semble pas soumise à l'ordre d'un plan. La TEI prévoit ce type d'agencement « lâche » avec l'élément `<dictScrap>`[35].

Le lecteur d'un dictionnaire moderne de langue française est plus habitué à un plan hiérarchique des articles. Deux étapes sont intervenues depuis Du Cange. Jonhson (1755) introduit la numérotation des emplois dans son *Dictio-*

[28] Par exemple pour la référence "E.*Hel.*1453", citée et balisée à l'identique dans le LSJ et le *DGE*, il s'agira de récupérer, via la balise <bibl> du LSJ Perseus, l'URI Perseus affichant le texte d'Euripide.

[29] http://www.stoa.org/sol/

[30] Toutes deux consultables depuis le site http://papyri.info/

[31] http://epigraphy.packhum.org/inscriptions/

[32] http://epidoc.sourceforge.net/index.shtml

[33] Des premiers comptages bruts arrivent à 400 000 formes différentes pour 6 000 000 d'occurrences.

[34] Bon, B. - Guerreau-Jalabert, A. (2002), p.74.

[35] « <dictScrap> (dictionary scrap) encloses a part of a dictionary entry in which other phrase-level dictionary elements are freely combined. »

nary of the English language. Féraud (1787) importe la liste numérotée dans la lexicographie française, Littré (1840) mena le procédé à sa limite, en cherchant pour chaque mot une raison linéaire qui expliquerait l'évolution de ses acceptions dans l'histoire.

Pour sa part, la lexicographie grecque suivait déjà un plan hiérarchique qui reposait sur l'évolution historique du sens des mots, au moins depuis Franz Passow, *Handwörterbuch der griechischen Sprache* (1831), repris par Liddell & Scott (1843). L'agencement hiérarchique que l'on peut observer dans le plan des articles du *DGE* est bien le fruit de cet héritage, mais celui-ci a été revu et adapté en fonction des importantes positions théoriques linguistiques qui sous-tendent le projet, centrées sur une conception distributionnaliste du sens.[36] Ainsi, la nécessité n'a pas été uniquement de fournir, par rapport à ses prédécesseurs, une plus grande précision des faits morphologiques et d'augmenter le nombre de lemmes et de citations, mais aussi d'améliorer le traitement sémantique du lexique étudié. Le *DGE* peut donc être considéré aussi comme le résultat d'un héritage plus récent, et une mise en pratique, pour le grec ancien, de la théorie saussurienne des dimensions syntagmatique et paradigmatiques de la signification. A été par conséquent abandonné le principe de classification chronologique ou logique. L'ambition première des rédacteurs du *DGE* est de dessiner la « carte sémantique »[37] de chaque mot grec, et « de poursuivre et de perfectionner l'organisation *ramifiée* des acceptions sur une base sémantique »[38].

Si le *Du Cange* n'est qu'une suite de paragraphes, chaque article du *DGE* est un arbre strict et validé de significations, des `<sense>` récursivement enchâssés.

4.2. Traductions : la langue cible comme principe d'organisation de la langue source, <trans>

Le *Glossarium* ne propose pas de traduction, au grand regret des débutants, mais peut-être au profit de leur intelligence. La traduction moderne produit beaucoup de contresens. Considérez par exemple le mot *sensualitas*. Sous l'entrée « SENSUALITÉ », Godefroy (1901) propose la glose : « l'ensemble de nos sens ». Le lexicographe semble dire que la *sensualité* médiévale recouvre exactement notre usage moderne du mot. Il donne cet exemple : « Et il soit ainsi que ledit Pierre depuis un an en ça, par impatience, fragilité ou diminution de son corps et de sa *sensualité*, soit devenu tout ydiote. (1376, Arch. JJ 110, pièce 208.) ». Traduire *sensualité* par *sensualité* rend cette phrase incompréhensible, ou curieusement visionnaire. Une psychologie contemporaine pourrait en effet soutenir que le manque de contact sensuel rend l'enfant autiste, *ydiot*, « qui a l'esprit très borné » ; mais est-ce que cette interprétation convient à un document juridique de 1376 ? Dans le *Glossarium*, Carpentier (1766) cite le même passage avec pour glose : « Sensus, intellectus ». La *sensualité*, entendue comme faculté de faire sens, permet de mieux imaginer ce qui est arrivé à ce Pierre. Le lecteur de ces documents médiévaux est continuellement confronté à ces risques de méprises, témoignant d'une langue, et d'une société, structurée avec d'autres catégories. En conservant la langue source comme métalangage, les érudits classiques ont évité le biais induit par la traduction systématique, inévitable avant la prise de conscience structurale.

Les méprises sur le grec ancien sont tout aussi probables, mais la communauté autour de cette langue est plus ancienne, plus importante, si bien que la responsabilité du lexicographe est moindre. Le chercheur consultera plus d'une référence, permettant au *DGE* de prendre un parti fort. Il a été le premier ouvrage de la lexicographie grecque non spécialisée construit systématiquement suivant un modèle sémantique componentiel, et la rédaction des articles est présidée par certains principes sémantiques fondamentaux :

- un mot n'a pas de signification autonome, ses significations sont en fonction de ses *distributions*, qui varient d'une langue à l'autre ;
- un même mot peut entrer dans des systèmes d'*opposition* divers à l'intérieur d'un ou plusieurs champs sémantiques ;
- les sèmes pertinents n'ont rien d'universel : chaque langue a les siens et les structure de manière *unique* (« anisomorphisme » des langues)

La théorie trouve son application dans la rédaction de ce dictionnaire bilingue. Les articles du *DGE* suivent une organisation en fonction de la langue cible, l'espagnol : « la traduction permet de regrouper des faits et de tracer des lignes d'organisation »[39]. Il en résulte une structure hiérarchique, « ramifiée », visant à perdre le moins possible les distinctions du grec. Les sens, `<sense>+`, sont structurés selon les traductions, `(trans(tr)+)`, qui restent à un niveau équivalent, ce qui est linguistiquement convenable. L'exemple suivant décrit un verbe dont les équivalents traductionnels proposés sont distribués à partir de patrons syntaxico-sémantiques :

διακλέπτω
<sense>**I**
<sense>**1** <usg>c. ac. de cosa</usg>
<trans><tr>*robar*</tr></trans>…

[36] Cette approche sémantique est aussi celle du dernier *Supplement* du LSJ (1996), et sera encore plus systématisée par le très intéressant *Greek Lexicon Project* actuellement mené à Cambridge,
http://www.classics.cam.ac.uk/faculty/research_groups_and_societies/greek_lexicon/

[37] « mapa semántico », Adrados - Gangutia - López Facal - Serrano Aybar (1977), p.265.

[38] Adrados - Rodríguez Somolinos (2005), p.290.

[39] Adrados - Gangutia - López Facal - Serrano Aybar (1977), p.267.

```
 </sense>
 <sense>2 <usg>c. ac. de pers.</usg>
  <trans><tr>salvar la vida</tr>, sustraera un peligro</tr></trans>…
 </sense>
 <sense>3 <usg>c. ac. de abstr.</usg>
  <trans><tr>eludir</tr>, <tr>esquivar</tr></trans>…
 </sense>
</sense>
<sense>II <usg>intr. en v. med.</usg>
 <trans><tr>esconderse</tr>, <tr>ocultarse</tr></trans>…
</sense>
```

Remarquons que la précision typographique, contrôlée semi-automatiquement par une utilisation avancée du traitement de texte, a grandement favorisé la précision du balisage : le champ de traduction, `<trans>` contient un ou plusieurs équivalents traductionnels en italique, `<tr>`, séparés par une virgule ou autre segment de texte (« o », « y », etc.) en caractères romains. Là encore, l'exploration de l'ouvrage en ligne y gagnera en exactitude et en finesse, et les chercheurs en linguistique computationnelle pourront exploiter les nuances que l'espagnol ouvre sur les mots grecs.

4.3. Marques : Étiquetage sémantique et schémas distributionnels, <usg>

Pour assurer une meilleure approche du système de la langue décrite, et en réduire le moins possible les distorsions inhérentes à la transposition dans une autre langue, les traductions du *DGE* sont très souvent complétées par des paraphrases, des gloses et autres types d'informations que nous qualifierons de marques d'usage. Les rédacteurs ont développé un effort très important pour les utiliser avec une plus grande systématicité que ses prédécesseurs. Leur encodage informatique peut alors contribuer aux recherches pour l'annotation sémantique des corpus de grec ancien. On sait que les dictionnaires peuvent en effet aider à établir des classes sémantiques et des schémas distributionnels[40]. Ces ressources linguistiques s'ajoutent à la lemmatisation pour explorer les corpus. Elles permettent de croiser des familles de mots, tout en servant de clefs de tri à des concordances plus étendues. Si, par exemple, on observe la structure de l'article διακλέπτω présenté ci-dessus, la tâche consiste à, d'une part, exploiter le vocabulaire espagnol des traductions pour attribuer à chaque lemme une classe sémantique (humain, objet, animal, dieu). Le *DGE* indique d'autre part de nombreux patrons de construction syntaxique, d'emplois dictés par la rection et il sera possible d'extraire la grille argumentale des verbes en repérant des motifs réguliers tels que: « c. (con, avec) + acc., dat., etc. (accusatif, datif, etc.), de pers., de cosa, etc. (de personne, de choses, etc.) », « intr. », « abs. ». Il est par exemple intéressant d'observer les glissements sémantiques à l'œuvre dans la combinaison du verbe διακλέπτω avec un complément à l'accusatif désignant une chose (*robar*, « voler »), un humain (« sauver la vie, soustraire à un danger »), une entité abstraite (*eludir*, *esquivar* « éluder, esquiver ») ou à la voix moyenne sans complément (*esconderse*, « se cacher »). De tels projets justifient en tous cas un balisage soigné du *DGE*, au delà de l'affichage typographique : ces informations peuvent être extraites et normalisées, et le dictionnaire peut aider à enrichir une base lexicale de traits sémantiques que l'ordinateur peut exploiter.

Il semble impossible de dégager formellement de tels éléments du flux discursif du texte du *Du Cange*. Cependant, le tome X comporte 45 *indices* sur 100 pages (CXVII-CCXVI). Ces listes de mots ont des titres comme « *Corpus ; corporis humani et animalium partes*. », « *Pisces, piscatura*. », ou « *Dignitates civiles, palatine, militares, honores, officia*, etc. » L'informatisation de ces indices est en cours. Ces classifications étrangement désordonnées présentent l'intérêt de provenir du lexique lui-même et non d'ontologies modernes. Les expériences en corpus jugeront de la pertinence de ces catégories.

5. Conclusion

La confrontation d'un glossaire ancien, le *Du Cange*, à un dictionnaire récent tel que le *DGE* montre, d'une part, jusqu'à quel point la lexicographie a structuré son texte en trois siècles, avançant aux franges de l'exactitude informatique. Elle montre d'autre part qu'une telle rencontre est rendue maintenant possible et pertinente grâce à l'utilisation partagée d'un même standard d'encodage (ici la TEI), fournissant un même vocabulaire de comparaison.

Outre la connaissance, cette grille permet aussi un programme d'action. Le balisage d'un dictionnaire imprimé peut ainsi s'approfondir en visant progressivement trois objectifs :

1. Présentation à l'écran — Elle peut commencer par les images, puis le texte, puis la typographie, et en tous cas, demande l'identification des lemmes, permettant la navigation dans les renvois.
2. Formulaire de recherche avancée — L'indexation de champs d'information demande que la typographie soit interprétée en fonctions lexicographiques, comme : citations, références, nature, domaines… L'apparence à l'écran supporte sans gêne des imprécisions qui peuvent être refusée au moteur de recherche. Une citation en italique non reconnue ne se remarque pas à l'oeil, par contre elle ne sortira jamais dans les résultats. La fiabilité à un coût en vérifications automatisées ou assistées.

[40] Sur ce sujet, voir par exemple Valette, Mathieu *et al.* (2006).

3. Ressources linguistiques — L'extraction de données automatiquement exploitables impose un degré d'exigence supplémentaire, au-delà de ce qui se voit et se cherche, car elles vérifient le dictionnaire en l'appliquant sur le corpus : les codes de nature et de flexion doivent être normalisés, les marques de domaine ne servent plus seulement à comprendre mais à classifier automatiquement, les traductions sont renversées risquant les incohérences, les indications syntaxiques ne sont plus seulement conçues mais testées…

La prise en compte de ces objectifs approfondit les exigences d'une lexicographie scientifique. L'informatique a jusqu'ici permit de gagner du temps, avec le traitement de textes simplifiant la typographie, ou les bases textuelles, mais elle ne modifiait pas le but : donner une page exacte à lire. Avec un dictionnaire conçu comme ressource linguistique automatisable, l'informatique n'est plus un accessoire utile, elle modifie le projet, et augmente la charge de travail sans qu'elle puisse être mesurée aussi clairement que par la nouvelle parution d'un fascicule.

Est finalement en jeu, dans l'informatisation de ces deux ouvrages, le renouvellement du rôle des données lexicographiques « traditionnelles » dans l'avancement qualitatif de l'exploration et l'annotation des corpus anciens : ils sont des sources de lemmes dont l'authenticité est garantie par l'analyse des lexicographes, des sources valables, voire parfois uniques dans le cas du *Du Cange*, de références vers des corpus, et des modèles de structuration du lexique en fonction de traits et de catégories syntaxiques et sémantiques exploitables.

6. Bibliographie

Adrados, F.R. - Gangutia, E. - López Facal, J. - Serrano Aybar, C. (1977), *Introducción a la lexicografía griega*, Madrid, CSIC.

[DGE] Adrados, F.R. dir. (1980-), *Diccionario Griego-Español*, Madrid, Consejo Superior de Investigaciones Científicas, 6 vol. parus + rééd. vol. I en 2008.

Adrados, F.R. - Rodríguez Somolinos, J., (eds.) (2005), *La lexicografía griega y el Diccionario Griego-Español*, *DGE. Anejo* VI, Madrid, CSIC.

Bon, B. - Guerreau-Jalabert, A. (2002), « Pietas : réflexions sur l'analyse sémantique et le traitement lexicographique d'un vocable médiéval », *Médiévales*, n° 42, p. 73-88. http://www.persee.fr/web/revues/home/prescript/article/medi_0751-2708_2002_num_21_42_1540

[Du Cange] du Cange, Charles du Fresne (sieur) et al. (1678-1887), *Glossarium mediae et infimae latinitatis*, éd. augm., Niort, L. Favre, 1883-1887, 10 vol. in-quarto.

Ide, N - Véronis J. (1996), « Codage TEI des dictionnaires électroniques », *Cahiers GUTenberg*, 24, Rennes, p.170-176.

[LMPG] Muñoz Delgado, L. (2001), *Léxico de magia y religión en los papiros griegos*, Madrid, Consejo Superior de Investigaciones Científicas.

[LSJ] Liddell, H.G. - Scott, R., *A Greek-English Lexicon*, [1e éd.1843], 9e éd. revue par Jones H.S. assisté de McKenzie R., 1940, *Revised Supplement*, Glare P.G.W. éd., Oxford, Clarendon Press, 1996.

Merrilees, B. (1996), *The Shape of the Medieval Dictionary Entry*, Toronto. <http://www.chass.utoronto.ca/epc/chwp/merrily2/>

[TEI] TEI Consortium (1999, 2002, 2007), *Guidelines for Electronic Text Encoding and Interchange*. Oxford — Providence — Charlottesville — Nancy, C.M. Sperberg-McQueen and Lou Burnard. http://www.tei-c.org/release/doc/tei-p5-doc/html/

Tutin, A.- Véronis, J. (1998), *Electronic Dictionary Encoding : Customizing the TEI Guidelines*, Eighth Euralex International Congress (EURALEX'98), Liège, p.4-8. http://www.up.univ-mrs.fr/veronis/pdf/1998euralex.pdf

Valette, Mathieu *et al.* (2006), « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens. », *Verbum ex machina, Actes de la 13ème conférence sur le traitement automatique des langues naturelles* (TALN 06). http://www.revue-texto.net/Corpus/Publications/Valette_Estacio.pdf

Wooldridge, R. (1977, 1997), Les Débuts de la lexicographie française : Estienne, Nicot, et le Thresor de la langue françoyse (1606).

http://www.chass.utoronto.ca/~wulfric/edicta/wooldridge/