# D'un dictionnaire de lemmatisation (*D.A.G.*) à un dictionnaire dérivationnel du grec ancien (*D.D.G.*)

Raphaël Gérard, Bastien Kindt

Université catholique de Louvain – Institut orientaliste – Place Blaise Pascal, 1
1348 Louvain-la-Neuve – Belgique
gerard@ori.ucl.ac.be, kindt@ori.ucl.ac.be

## Abstract

The lemmas of the *Dictionnaire Automatique Grec* (*D.A.G.*), developed at the U.C.L. for the lemmatisation of patristic and Byzantine Greek texts, are subject of tagging of the constituting morphemes. These data are gathered in a *Dictionnaire Dérivationnel Grec* (*D.D.G.*). By means of specific interfaces, lexical tools make a selection of lemmas with common morphemes possible; a listing of all words – composed or derived – belonging to the same theme may also be visualised as a morphological tree.

## Résumé

Les lemmes du *Dictionnaire Automatique Grec* (*D.A.G.*), développé à l'U.C.L. pour la lemmatisation des sources grecques patristiques et byzantines, font l'objet d'un étiquetage systématique des morphèmes qui les constituent. Ces données sont rassemblées dans le *Dictionnaire Dérivationnel Grec* (*D.D.G.*). Une interface d'interrogation permet de sélectionner les lemmes partageant des morphèmes communs. Une autre permet d'afficher sous la forme d'un arbre dérivationnel tous les mots, composés ou dérivés, issus d'un même thème.

**Mots-clés :** dictionnaire électronique, dictionnaire dérivationnel, étiquetage, grec ancien, lemme, lexique, morphème, récursivité

## 1. Introduction

### 1.1. *Le Dictionnaire Automatique Grec (D.A.G.) : présentation*

Le *D.A.G.* est un dictionnaire électronique de la langue grecque ancienne élaboré par deux équipes de l'U.C.L., celle du « Projet de recherche en lexicologie grecque » de l'Institut orientaliste[1], et celle du Centre de Traitement Automatique du Langage (CENTAL)[2]. Utilisé comme dictionnaire de référence pour la lemmatisation automatisée des sources grecques patristiques et byzantines[3], sa nomenclature, structurée en une base de données relationnelle, rassemble les matériaux lexicaux dénombrés au fil des traitements successifs. En tenant compte des analyses passées et en cours, l'ensemble porte sur un *corpus* totalisant 4.284.380 mots-occurrences ; 174.758 formes différentes constituent la microstructure du dictionnaire,

[1] La bibliographie complète du projet est disponible sur la toile à l'adresse http://tpg.fltr.ucl.ac.be ; cf. aussi trois contributions récentes : Coulie (2003), Kindt (2003a) et Kindt (2003b).

[2] L'ancien Centre de Traitement Électronique des Documents (CETEDOC), cf. http://cental.fltr.ucl.ac.be.

[3] Ce choix thématique découle des domaines de recherches propres à l'Institut orientaliste. Les concordances sont publiées dans le *Thesaurus Patrum Graecorum* (*T.P.G.*), une sous-collection du *Corpus Christianorum* diffusée par Brepols Publisher, cf. http://www.brepols.net et http://www.corpuschristianorum.org, ainsi que le site du *T.P.G.* cité note i.

33.874 lemmes, sa macrostructure (cf. fig. 3). Les analyses effectuées portent majoritairement, mais non exclusivement, sur des sources patristiques du IV[e] s. ap. J.-C. La nature fortement classicisante de la langue de ces textes, ainsi que des principes explicites de formulation des lemmes et la stabilité des normes de dépouillement des formes, rendent l'utilisation du dictionnaire opérationnelle sur des sources antérieures, d'époque classique en l'occurrence, ou postérieures[4]. Depuis 1987[5], il a subi d'importantes modifications susceptibles de lui conférer, progressivement, les potentialités caractéristiques des outils contemporains d'analyse lexicale.

Dans le cadre de la lemmatisation, la structure du *D.A.G.* a été rendue comparable à celle des dictionnaires du Laboratoire d'Automatique Documentaire et Linguistique (L.A.D.L.) de l'Université de Marne-la-Vallée (Paris)[6]. Une étiquette relative à la classe morpho-syntaxique a de plus été attachée aux lemmes. Les textes sont traités sous le logiciel UNITEX[7]. Des grammaires locales de désambiguïsation y ont été adaptées ; la levée des ambiguïtés lexicales, étape laborieuse de la lemmatisation, sera désormais partiellement automatisée (Kindt, 2003b : 10-12).

En dehors du cadre de la lemmatisation, les modes d'interrogation du dictionnaire ont été multipliés grâce à un encodage systématique des morphèmes constitutifs des lemmes. La présente contribution détermine la nature précise d'une telle opération, en présente les objectifs, et décrit les applications informatiques créées à cet effet.

### 1.2. *L'étiquetage des morphèmes constitutifs des lemmes du* D.A.G. *: nature et objectifs*

La lemmatisation consiste en un étiquetage lexical basé sur l'unité-mot. Les champs « lemmes » et « formes » de la base de données du *D.A.G.* peuvent faire l'objet d'interrogations. La formulation de ces requêtes se réduit toutefois à une chaîne de caractères, ce qui peut générer des réponses inadéquates. Une recherche sur <χειρ> « main », fournira ἀκροχειρίζω « toucher du bout des doigts », ἐπιχείρησις « entreprise », ou χειροποιητός « fait de main d'homme », réponses adéquates, mais aussi χείρων « pire », sans rapport avec χείρ « main ». Pour rendre l'exploration du *D.A.G.* plus efficace, il faut réduire le « bruit » amené par le manque de précision de ces requêtes. L'étiquetage lexical organisé au départ de l'unité-mot a donc été complété par un étiquetage des unités inférieures au mot, les morphèmes, et même, pour prendre en compte les thèmes, les combinaisons de morphèmes.

Une telle démarche requiert deux opérations concomitantes : le dénombrement des morphèmes et la mise en relation des lemmes présentant des morphèmes communs. Une base de données originale a donc été developpée. Au départ de celle-ci, une interrogation basée sur χείρ « main », fournit les septante-trois lemmes apparentés à ce mot (de ἀκρόχειρ, ἀκροχειρίζω, ἀνεπιχείρητος, ἀντίχειρ, ἀντιχειροτονέω, ἀπαρεγχείρητος, etc., à χειρώναξ). De plus, comme la nature des morphèmes et des combinaisons de morphèmes est annotée, l'interrogation peut être complétée par une requête touchant n'importe quel élément constitutif des mots, le suffixe -ιζε/ο-, par exemple, formant les verbes dénominatifs en -ίζ-ω/-ο-μαι-. La réponse affiche alors cent douze lemmes (depuis ἀκροχειρίζω, ἐγχειρίζω, ἀναλογίζομαι, διαλογίζομαι, etc., jusqu'à ὑπνίζω) (Gérard et Kindt, 2003). Ces matériaux et les interfaces

[4] Sur toutes ces notions, cf. Kindt (2003a : 1-19).

[5] Date de la première mention du *D.A.G.* dans Denis (1987 : XII).

[6] Le *D.A.G.* devient ainsi un DAG_DELAF ; sur les dictionnaires DELA, cf. Courtois (1990 : 11-22). Cf. aussi le site du Laboratoire d'Informatique Linguistique de l'Université de Marne-la-Vallée à l'adresse http://infolingu.univ-mlv.fr.

[7] Sur UNITEX, cf. http://www-igm.univ-mlv.fr.

de saisie et d'interrogation qui permettent de les manipuler sont rassemblés dans une nouvelle base de données baptisée *Dictionnaire Dérivationnel Grec* (*D.D.G.*). À ce stade, l'étiquetage est manuel : le travail progresse lemme après lemme, les grandes familles de mots appartenant à un même champ dérivationnel étant privilégiées. À l'avenir, l'analyse semi-automatique des constituants morphologiques des lemmes nouveaux est envisageable.

### 1.3. *Une idée ancienne, des moyens nouveaux*

Le *D.D.G.* est un dictionnaire affranchi de l'ordre alphabétique. Malgré son intérêt pratique, ce mode de classement traditionnel n'a pourtant aucun fondement linguistique. Il présente de plus un effet pervers qui est d'éloigner les uns des autres un grand nombre de mots linguistiquement apparentés. Regrouper les mots issus d'un même champ dérivationnel n'est pas une idée neuve. Dans son article intitulé *Towards an Electronic Greek Historical Lexicon* (1994), W.A. Johnson « rêve » d'une telle réalisation que les moyens informatiques contemporains rendaient déjà envisageable (Johnson, 1994). Prenant pour exemple le nom μανία et les adverbes en -κως, il illustre l'intérêt que présenterait un inventaire global des mots dérivés d'une même base et mis en relation avec la chronologie des sources dans lesquelles ces créations lexicales apparaissent pour la première fois. Il cite également le *Thesaurus Graecae Linguae* de H. Estienne, car la nomenclature de l'édition originale de ce monument de la lexicographie (1572) était structurée selon les racines dont les mots grecs dérivent, ἐπεισβαίνω figurant ainsi à la suite d'autres dérivés et composés de βαίνω, et non à la place que lui aurait assignée le respect de l'ordre alphabétique *stricto sensu* [8].

Trois facteurs ont facilité l'opération d'étiquetage des morphèmes. Le premier tient au fait que le *D.A.G.* était déjà constitué ; l'ensemble de ses lemmes propose un échantillon, incomplet, certes, au regard de la totalité du lexique grec, mais déjà représentatif. Le deuxième tient aux morphèmes qui relèvent, quant à eux, d'une série limitée d'éléments identifiables et distinctifs organisés en système, un ensemble d'éléments « nécessairement présents »[9] dans l'échantillon de référence. Le dernier facteur tient à l'essor de l'outil informatique ; une organisation appropiée de la base de données permet de fusionner en une seule étape la saisie des éléments morphologiques pertinents et leur mise en relation avec les lemmes dans lesquels ils apparaissent. Le travail présenté ici dépasse le rêve de W.A. Johnson car il suffira, à l'avenir, d'appliquer le *D.D.G.* à un *corpus* textuel pour dégager les procédés de formations lexicales qui s'y actualisent. Il dépasse aussi la réalisation de H. Estienne car il est désormais possible de classer les mots grecs non seulement selon leur racine, leur radical ou leur thème, mais aussi selon chacun des morphèmes (ou des combinaisons de morphèmes) qui les constituent.

## 2. Le *D.D.G.*

### 2.1. *Structure*

Les matériaux lexicaux présents dans le *D.D.G.* sont, à l'instar de ceux du *D.A.G.*, structurés en une base de données relationnelle. Celle-ci renferme plusieurs tables. La majorité de celles-ci rassemble les données morphologiques permettant de décrire les lemmes, c'est-à-dire les morphèmes *proprie dictu* (radicaux, racines indo-européennes, suffixes, préfixes, préverbes), et les combinaisons de morphèmes (thèmes). Une table rappelle les lemmes du *D.A.G.*

[8] H. Estienne s'explique lui-même sur la conception du *Thesaurus Graecae Linguae*, cf. l'édition récente des préfaces du lexicographe dans Kecskeméti *et al.* (2003), et spécialement, pour le regroupement des mots par racine, pp. 250-251 et 293-294.

[9] Expression reprise à D. et P. Corbin dans Corbin et Corbin (1991 : 147).

Une dernière table établit les relations entre les lemmes et les morphèmes. Une interface appelée *Interface de Caractérisation Morphologique des Lemmes* (*I.C.M.L.*) permet d'encoder ou d'afficher les données du *D.D.G.* Deux interfaces d'interrogation permettent d'en explorer le contenu.

### 2.2. *Formulaire d'étiquetage des morphèmes*

Le formulaire *I.C.M.L.* est utilisé pour identifier et encoder les morphèmes et les combinaisons de morphèmes des lemmes choisis par l'utilisateur, qu'il sagisse d'un mot simple, d'un dérivé ou d'un composé.

Pour le dérivé ἐπανάβασις, par exemple (cf. fig. 1), les champs suivants sont affichés à l'écran : – le lemme en cours (ἐπανάβασις) ; – sa base dérivationnelle (ἐπαναβαίνε/ο-) ; – son thème (ἐπαναβασι-) ; – le suffixe attaché au thème (-τ/σι- (δόσις)) ; – la désinence du lemme (-ς) ; – la nature du lemme (nom commun). Les champs « radical » et « racine » restent vides car ils ont déjà été saisis dans le formulaire correspondant au lemme simple βαίνω, respectivement βα(ν) et *$g^w$em-/$g^w$e$h_2$. Ces deux informations, communes à plusieurs lemmes apparentés, ne sont donc encodées que sous le formulaire du mot simple. La structure même de la base de données permet ensuite, lors d'une interrogation, de répercuter ses informations sur tous les lemmes concernés.

*Figure 1. Formulaire d'étiquetage des morphèmes pour le lemme ἐπανάβασις.*

Pour le composé φιλάγαθος, le formulaire indique, outre le lemme : – son thème (φιλα-γαθε/ο-) ; – le premier élément de composé (φιλε/ο-) ; – le second élément de composé (ἀγαθε/ο-) ; – la désinence du lemme (-ς) ; – la nature du lemme : adjectif.

Certaines entrées du *D.D.G.* sont précédées de l'arobase (par exemple, @ἀκρόβατος) indiquant que le mot est bien attesté dans les sources — ce qui établit son existence dans la langue —, mais qu'il ne figure pas dans la nomenclature du *D.A.G.* et est postulé afin d'expliquer les lemmes qui en dérivent (tel ἀκροβατέω). L'astérisque marque les entrées du *D.D.G.* qui, quoique absentes du *D.A.G.* et des sources disponibles, doivent être postulées au niveau de la langue pour justifier d'autres formations lexicales qui en dérivent et qui, elles, sont présentes dans le *D.A.G.* (par ex. *σκληροφάγος nécessaire pour décrire le lemme σκληροφαγία).

D'autres boutons du formulaire permettent d'appliquer des filtres, d'activer des raccourcis vers le *D.A.G.* lui-même, vers les interfaces d'interrogation et vers différents outils lexicographiques.

### 2.3. *Formulaire d'étiquetage de la classe morpho-syntaxique des lemmes*

Les lemmes ont reçu une étiquette identifiant leur classe morpho-syntaxique (cf. fig. 3). Les classes retenues sont les suivantes : nom commun ; nom propre (anthroponymique, patronymique, toponymique, ethnique) ; adjectif ; verbe ; pronom ; article ; muméraux (cardinaux, ordinaux) ; invariable (adverbe, conjonction, interjection, négation, préposition).

### 2.4. *Interfaces d'interrogation*

#### 2.4.1. *Interface de recherche à critères multiples*

L'« interface de recherche à critères multiples » sélectionne des éléments et fournit les lemmes dans lesquels ceux-ci s'actualisent. L'utilisateur détermine d'abord le type d'élément (thème, suffixe, préverbe, préfixe ou radical). Il sélectionne ensuite un des éléments proposés dans une listbox et lance la requête. Le résultat s'affiche sous forme de liste. Quand une recherche est basée sur un thème, une zone permet de préciser s'il s'agit d'une base dérivationnelle, d'un premier ou d'un second élément de composé. Une autre zone permet de sélectionner des « critères supplémentaires » : la classe morpho-syntaxique des lemmes attendus par l'utilisateur et la nature de la base dérivationnelle, du premier ou du second élément de composé. Il est ainsi possible de rechercher quels sont les lemmes adjectivaux (« critères supplémentaires ») construits sur le thème (« type d'élément ») βαινε/ο- (« élément » tiré de la liste des thèmes). La réponse fournit les lemmes ἄβατος, βατός, δύσβατος et @ἀκρόβατος. La recherche peut porter simultanément sur trois « types d'élement » reliés par des opérateurs booléens. Il est ainsi possible d'extraire du *D.D.G.* les lemmes construits sur le thème βαινε/ο- et comprenant le suffixe -τ/σι- (δόσις) : le mot βάσις s'affiche. En excluant le suffixe -τ/σι- (δόσις), dix-sept réponses s'affichent, de βῆμα et ἄβατος à βάθρα. En limitant la requête aux lemmes dans lesquels le thème βαινε/ο- apparaît comme second élément de composé, dix résultats apparaissent : ἄβατος, δύσβατος, etc. En sélectionnant le thème βαινε/ο- et le suffixe -μα(τ)-, le résultat est βῆμα.

#### 2.4.2. *Interface de recherche par descendance*

L'« Interface de recherche par descendance » fait état du caractère récursif des mots grecs. La récursivité est la propriété par laquelle le lexique renouvelle son stock d'unités-mots par dérivations ou compositions successives. Une base dérivationnelle accrue d'un suffixe est ainsi à

l'origine d'une création lexicale qui, à son tour, sera susceptible de constituer une nouvelle base : βαίνω → καταβαίνω → συγκαταβαίνω → συγκατάβασις.

L'« Interface de recherche par descendance » permet à l'utilisateur de sélectionner une base dérivationnelle et lui fournit tous les lemmes qui y sont apparentés. Le résultat s'affiche sous la forme d'une « arborescence dérivationnelle » faisant état de la récursivité du thème retenu (cf. fig. 2).

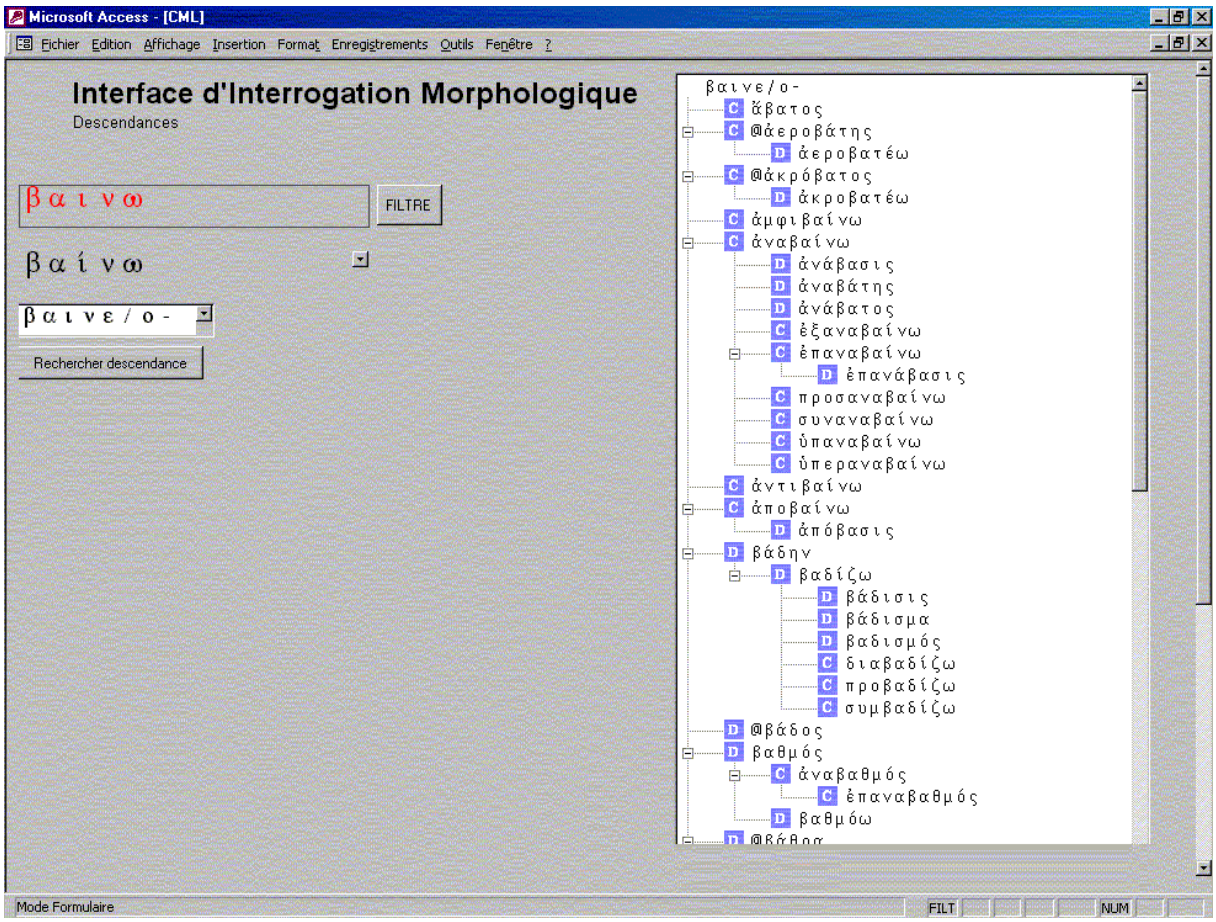


*Figure 2. Arborescence dérivationnelle de βαίνω (extrait).*

## 4. Conclusions

### 4.1. *État d'avancement du projet*

Le travail d'étiquetage des morphèmes s'étend à ce jour à 10.424 lemmes (soit 30,77% des matériaux lexicaux du *D.A.G.*). La classe morpho-syntaxique de 31.844 lemmes a été définie (cf. fig. 3). Ces tâches doivent se poursuivre. Dans un premier temps, les noms propres, les pronoms, l'article, les numéraux et les invariables ont été arbitrairement écartés de la caractérisation morphologique (ils représentent 5.430 lemmes, soit 16% du *D.A.G.*). Les anthroponymes grecs seront prioritairement pris en compte dans les analyses futures.

Ces données doivent être éprouvées et contôlées. Les interfaces d'interrogation permettront par exemple de vérifier la pertinence du dégroupement de certains suffixes homographes, tels -ιζε/ο- suffixe déverbatif (par exemple dans βαπτίζω, dont la base dérivationnelle est le verbe βάπτω), et -ιζε/ο- suffixe dénominatif (par exemple dans οἰκίζω, dont la base dérivationnelle

est le nom οἶκος). Les mêmes outils seront utilisés pour vérifier l'homogénéité des traitements. L'analyse des formations dites parasynthétiques sera par exemple réexaminée. Dans le *D.D.G.*, le lemme ἀδίδακτος dérive directement de διδάσκω et non de διδακτός. Est-ce que la même interprétation a été adoptée pour les autres formations analogues ?

| | | | |
|---|---|---|---|
| ***D.A.G.*** 4.284.380 mots-occurrences | | | |
| formes | 174.758 | lemmes | 33.874 |
| ***D.D.G. (étiquetage des morphèmes)*** 10.424 lemmes traités | | | |
| racines | 378 | préfixes | 10 |
| radicaux | 631 | préverbes | 12 |
| suffixes | 286 | thèmes | 10.779 |
| ***D.D.G. (étiquetage des classes morpho-syntaxiques)*** 31.844 lemmes traités | | | |
| noms | 12.740 | articles | 1 |
| noms communs | 8.972 | numéraux | 803 |
| noms propres | 3.762 | cardinaux | 703 |
| anthroponymiques | 1.842 | ordinaux | 100 |
| patronymiques | 17 | invariables | 823 |
| ethniques | 28 | adverbes | 645 |
| toponymiques | 1.704 | conjonctions | 47 |
| adjectifs | 8.730 | négations | 2 |
| verbes | 9.098 | interjections | 17 |
| pronoms | 43 | prépositions | 36 |

*Figure 3. Données chiffrées des matériaux lexicaux du D.A.G. et du D.D.G.*

### 4.2. *Perspectives ouvertes*

Tels qu'ils sont actuellement conçus, le *D.A.G.* et le *D.D.G.* permettent d'assurer la lemmatisation des sources et d'étudier les champs dérivationnels des mots enregistrés dans leur nomenclature commune. L'intérêt direct du *D.D.G.* est double. Lors de l'exercice de la lemmatisation, il offre les paradigmes utiles à l'étude de la productivité des morphèmes et de la régularité constructionnelle des mots rares rencontrés dans les sources. L'arborescence dérivationnelle permet d'établir facilement une liste, encore partielle, certes, mais fiable, des mots apparentés à un même terme, évitant aux utilisateurs un dépouillement fastidieux, et forcément imparfait, des dictionnaires traditionnels.

Les perspectives d'avenir pourraient orienter le projet vers la mise au point de programmes d'analyse. Un lemme inédit rencontré dans les sources ferait ainsi l'objet d'une caractérisation automatique de ses constituants. Les bases de données pourraient aussi accueillir des informations sémantiques permettant d'établir, par exemple, les relations de synonymie, d'antonymie, d'hyperonymie ou d'hyponymie qui traversent le lexique. Les applications présentes relient πύρ, ἐμπύριος et ζώπυρος, etc. Les outils futurs devraient rassembler ἐμπύριος « qui est en feu », αἶθοψ « qui est couleur de feu » et φλόγινος « enflammé, qui est couleur de feu ».

Enfin, les ressources et les outils conçus dans le cadre du projet seront progressivement mis à la disposition des utilisateurs sur la toile, *via* des interfaces web déjà opérationnelles en version expérimentale.

## Références

Coulie B. (1996). La lemmatisation des textes grecs et byzantins : une approche particulière de la langue et des auteurs. *Byzantion*, vol. (66) : 35-54.

Coulie B. (2003). Thesaurus Patrum Graecorum. In Leemans J. (Ed.), *Corpus Christianorum 1953-2003. Xenium Natalicum. Fifty Years of Scholarly Editing* : 169-172.

Corbin D. et Corbin P. (1991). Vers le Dictionnaire dérivationnel du français. *Lexique*, vol. (10) : 147-161.

Courtois B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, vol. (87) : 11-22.

Denis A.-M. (1987). *Concordance des Pseudépigraphes d'Ancien Testament. Concordance. Corpus des textes. Indices*.

Gérard R. et Kindt B. (2003). Le Projet de recherche en lexicologie grecque à l'Institut orientaliste de l'Université catholique de Louvain. La collection du Thesaurus Patrum Graecorum et le Dictionnaire Automatique Grec. *Byzantion*, vol. (73) [à paraître].

Johnson W.A. (1994). Towards an Electronic Greek Historical Lexicon. *Emerita*, vol. (62) : 253-261.

Kecskeméti J., Boudou B. et Cazes H. (2003). *La France des humanistes. Henri II Estienne, éditeur et écrivain* (*Europa Humanistica*).

Kindt B. (2003a). *La lemmatisation automatisée au service d'une description lexicale du grec ancien. Propositions pour la formulation des lemmes du* Dictionnaire Automatique Grec *(D.A.G.). Rapport de recherche présenté en vue de l'obtention du D.E.A. en Philosophie et Lettres (ISLE 3DA)*, orientation « Philologie et histoire orientales ». Louvain-la-Neuve.

Kindt B. (2003b). Avancées dans le traitement automatique du grec ancien à l'U.C.L. L'analyse des textes au service d'une description lexicale de la langue. Une description lexicale de la langue au service de l'analyse des textes. In Labbé D. (Éd.), *Lexicometrica*, numéro spécial « Autour de la lemmatisation » : 1-17 (http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema1.htm).