La lemmatisation automatisée des sources en grec ancien: Présentation de ressources linguistiques et d'outils de traitement

Sara Deodati
Libera Università "Maria SS. Assunta"
Facoltà di Lettere e Filosofia
Istituto di Linguistica
Via della Traspontina, 21
00 193 Roma – Italia
s.deodati@lumsa.it

Bastien Kindt
Université catholique de Louvain
Faculté de Philosophie et Lettres
Institut orientaliste
Place Blaise Pascal, 1
B 1348 Louvain-la-Neuve – Belgique
kindt@ori.ucl.ac.be

Abstract
These notes aim to describe the treatment applied to ancient Greek texts to standardise the forms indicated by accents of enclisis, ellipsis and contraction (crasis), identified and listed in various resources of UNITEX, a software conceived for the lexical and syntactic analysis of texts. All the applications illustrated concern Clemens of Alexandria's *Exhortation To The Greeks* and show how the developments in computer technology of the Research Project in Greek Lexicology advance towards the conception of a functional tool for analysis and a powerful research engine into the domain of ancient Greek language studies. The project will be fulfilled by the publication of lemmatised concordances. Lexical data supplied by this analysis are stored in an electronic dictionary (now resulting in 280,733 forms, classified under 58,598 entries) and in various linguitic resources.

1 Introduction

Toute entreprise de lemmatisation se heurte à deux difficultés: (i) la richesse morphologique de la langue traitée; (ii) l'ambiguïté. Les lignes qui suivent illustrent les réponses apportées à la première de ces difficultés dans le cadre d'un traitement automatique des textes en grec ancien (GA). Ces travaux s'inscrivent dans un projet d'analyse des sources patristiques et historiographiques d'époque byzantine (IVe-Xe s.) et se concrétisent par la production de concordances lemmatisées publiées dans le "Thesaurus Patrum Graecorum" (TPG)

(Coulie 1996, 2003).[1] Les matériaux lexicaux issus de ces analyses sont rassemblés dans un dictionnaire électronique, le "Dictionnaire Automatique Grec" (DAG) (Kindt 2004).[2] La formulation des lemmes répond à des normes explicites et stables assurant une description lexicale homogène des textes. Les traitements automatiques sont effectués sous une version adaptée de l'analyseur UNITEX.

Le recours au DAG assure une couverture des multiples variations formelles caractérisant les mots du GA. Il n'offre cependant pas une représentation lexicographique satisfaisante des variations accentuelles, des élisions et des contractions de formes propres à cette langue. L'objectif de cette contribution est donc triple: (i) présenter la version d'UNITEX utilisée (§2); (ii) présenter le traitement réservé aux formes marquées d'une variation accentuelle (§3.1), aux formes élidées (§3.2) et aux formes contractées (§3.3); (iii) présenter l'interface de lemmatisation (§4).[3] Ces différents points sont illustrés d'exemples tirés du *Protreptique* de Clément d'Alexandrie (ca 150-215 ap. J.-C.) (Deodati 2005), auteur dont le corpus complet est en cours de lemmatisation.[4]

2 Présentation de la version d'UNITEX adaptée au traitement du GA

UNITEX est un logiciel[5] conçu pour l'analyse lexicale et syntaxique des textes (Paumier 2003, 2004). Supportant le standard d'encodage UNICODE, il présente un environnement de travail indépendant de la langue et des alphabets utilisés. Son fonctionnement repose sur le recours à des ressources linguistiques externes: des fichiers de paramétrage de l'alphabet de la langue traitée (le fichier Alphabet.txt qui déclare, pour le GA, 269 couples d'équivalence entre les lettres, par ex. E = ἐ, ὸ = ό; le fichier Alphabet_sort.txt qui gère le tri alphabétique; Paumier 2004:140, Kindt et al. 2006), des dictionnaires électroniques (en l'occurrence le DAG) et des grammaires locales. Un premier traitement du texte (reposant sur l'application de grammaires) permet de le segmenter en phrases et de normaliser les graphies divergentes d'un même mot (une forme élidée est par exemple remplacée par sa forme complète corres-

[1] La direction de ce projet est assumée par le Professeur B. Coulie au sein de l'ARC "Diffusion des textes et des idées dans l'Orient chrétien" 01/06-266 (http://tpg.fltr.ucl.ac.be et http://nazianzos.fltr.ucl.ac.be). Le TPG est une sous-collection du "Corpus Christianorum" diffusé par Brepols Publishers (http://www.brepols.net). Les auteurs tiennent à exprimer leur gratitude envers Mlle A. Yannacopoulou pour la relecture attentive qu'elle a réservée à ce texte et pour ses nombreuses remarques et suggestions.

[2] 280.733 formes classées sous 58.596 lemmes, toutes classes morpho-syntaxiques sont représentées.

[3] Les développements informatiques sont assurés en collaboration avec le CENTAL (Centre de Traitement Automatique du Language; http://cental.ucl.ac.be) et le Laboratoire d'Informatique de l'Institut Gaspard Monge (Université de Marne-la-Vallée; http://infolingu.univ-mlv.fr). Pour l'ambiguïté, cfr. Kevers et al. 2005. La version standard d'UNITEX (1.2 beta, 20 mars 2006) propose un module de levée automatique des ambiguïtés lexicales nommé ELAG (Elimination of lexical ambiguities by grammars; Laporte et al. 1998-1999, Laporte 2001). Cet outil est déjà utilisé pour le GA mais n'est pas encore intégré à la version d'UNITEX adaptée au GA.

[4] B. Kindt est en charge de l'adaptation des ressources d'UNITEX au traitement du grec ancien. S. Deodati assume la lemmatisation du corpus de Clément d'Alexandrie (265.085 occurrences) et éprouve sur cet ensemble textuel les développements présentés dans cet article.

[5] Téléchargeable à l'adresse http://www-igm.univ-mlv.fr/~Unitex/download.html.

pondante). Ce prétraitement fournit un nouvel état du texte sur lequel est appliqué le dictionnaire. La liste et les effectifs des "mots occurrences" (words), des "formes de mots" (forms) et des "formes non reconnues" (unknown simple words) sont alors produits et consultables.

L'utilisateur peut ensuite rechercher des "motifs" et en afficher la concordance, soit dans UNITEX, soit dans un navigateur WEB. Les "motifs" peuvent être des "formes de mots" ou des "lemmes", mode d'interrogation classique, mais aussi des codes relatifs à la catégorie morpho-syntaxique des mots, des filtres morphologiques (Paumier 2004: 48-49) ou des combinaisons de ces différents éléments. La Figure 1 présente un extrait de la concordance répondant au motif de recherche <I+Prep><E+DET><N+Ant><<ης|ου|οιο|έος$>> permettant d'extraire du texte toutes les séquences constituées d'une préposition initiale, d'un article facultatif et d'un anthroponyme à finale en -ης, -ου, -οιο ou -έος. L'intégration des informations flexionnelles, indispensables pour une automatisation efficace des processus de levée des ambiguïtés, est en cours d'implémentation.[6]

2-39-6 διὰ τὴν συνήθειαν, Θηβαῖοι δὲ τὰς γαλᾶς **διὰ τὴν Ἡρακλέους** γένεσιν.{S}
2-34-3 τὸν βίον.{S} Διόνυσος γὰρ κατελθὼν **εἰς Ἅιδου** γλιχόμενος ἠγνόει τὴν ὁδόν
4-59-1 εὐστεφάνου τ' Ἀφροδίτης·{S} ὡς τὰ πρῶτα μίγησαν **ἐν Ἡφαίστοιο** δόμοισι λάθρῃ·
3-45-4 ἐν τῷ α' τῶν περὶ τὸν Φιλοπάτορα ἐν Πάφῳ λέγει **ἐν τῷ τῆς Ἀφροδίτης** ἱερῷ Κινύραν

Figure 1. Extrait de la concordance des séquences <I+Prep><E+DET><N+Ant><<ης|ου|οιο|έος$>>

UNITEX offre également une image graphique des textes. La Figure 2 présente le graphe de la phrase "Ὅση γε ἡ δύναμις τοῦ θεοῦ".

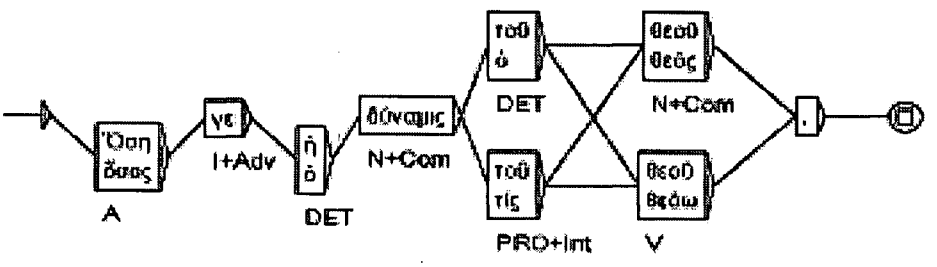

Figure 2. Graphe de la phrase

Chaque mot est représenté par une "boîte" mentionnant la forme, le lemme et son étiquette morpho-syntaxique. Les liens entre les boîtes symbolisent le continuum de la phrase. La forme θεοῦ est représentée par deux boîtes correspondant aux deux analyses proposées par le

[6] Une partie des formes du DAG est déjà dotée d'informations flexionnelles; cfr. note 3 et Kindt et al. 2006.

dictionnaire, respectivement un nom (θεός) et un verbe (θεόμ). Cette représentation visualise de manière commode les ambiguïtés (Paumier 2004: 146-147).

3 En amont de la lemmatisation

Les variations accentuelles, les élisions et les contractions sont traitées soit par un paramétrage correct du fichier Alphabet.txt soit lors de la phase de normalisation du texte.

3.1 *Reconnaissance des formes marquées d'un baryton ou d'un accent d'enclise*

En GA, l'accent d'un mot peut varier selon sa position ou son environnement dans la phrase: (i) un mot oxyton (accent aigu sur la dernière syllabe, par ex. φωτισμόν) devient baryton (accent grave, par ex. φωτισμὸν); (ii) précédé d'un enclitique, un mot proparoxyton (accent aigu sur l'antépénultième, par ex. ἄνθρωπος) reçoit sur la syllabe finale un accent supplémentaire dit "accent d'enclise" (ἄνθρωπός).[7] Les formes φωτισμὸν et ἄνθρωπός ne sont pas versées au dictionnaire. Mais, puisque le fichier de l'alphabet contient les correspondances ὸ = ό et ό = o, le système est à même de les assimiler aux entrées canoniques du DAG, φωτισμόν et ἄνθρωπος.

3.2 *Normalisation des formes élidées*

Les élisions apparaissent à l'initiale (᾽γαθέ pour ἀγαθέ), en finale (ἀφ᾽ pour ἀπό) ou à l'initiale et en finale (᾽πιοῦς᾽ pour ἀπιοῦσα).[8] Ces formes ne sont pas versées au dictionnaire. Une grammaire constituée de six graphes[9] permet de remplacer une forme élidée par sa forme complète correspondante (Figure 3).

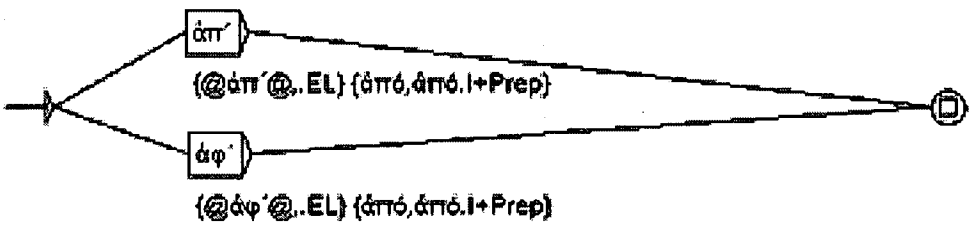

Figure 3. Extrait du graphe décrivant les élisions

Les formes ἀπ᾽ ou ἀφ᾽ sont ainsi remlacées par les séquences {@ἀπ᾽@,.EL} {ἀπό, ἀπό.I+Prep} et {@ἀφ᾽@,.EL} {ἀπό, ἀπό.I+Prep}, respetivement. Le premier ensemble entre accolades permet l'affichage dans la concordance de la forme élidée du texte, encadrée

[7] L'exemple fourni suffira pour illustrer le traitement proposé, même si la réalité de la langue est plus complexe.
[8] Dans le vers 651 d'*Iphigénie à Aulis* d'Euripide, cfr. Verraghenne 2005.
[9] En fait de grammaires il s'agit de transducteurs qui, appliqués au texte, recherchent des expressions (les éléments encadrés) et les remplacent par les sorties requises (les éléments entre accolades); cfr. Paumier 2004: 55.

d'arobases (Figure 4). Le code EL (pour "élision") peut intervenir dans un motif de recherche et permet d'afficher la concordance de toutes les formes élidées d'un corpus (Figure 4). Le second ensemble entre accolades fournit la forme résolue (ἀπό), son lemme (ἀπό) et sa catégorie morpho-syntaxique (I+Prep = préposition). Les informations contenues dans ces graphes proviennent des analyses antérieures. Dans leur état actuel, elles identifient et remplacent 254 formes élidées.

2-33-1 @δ'@ δὲ ἄρα χαῖται ἐπερρώσαντο ἄνακτος κρατὸς @ἀπ'@ ἀπὸ ἀθανάτοιο {S}
2-20-2 ποιμὴν δὲ ὁ Εὔμολπος, συβώτης δὲ ὁ Εὐβουλεύς· {S} @ἀφ'@ ἀπὸ ὧν τὸ Εὐμολπιδῶν

Figure 4. Extrait de la concordance des formes élidées basée sur la requête <EL>

3.3 *Normalisation des formes contractées*

La contraction de la finale vocalique d'un premier mot avec l'initiale vocalique d'un second mot produit une unité graphique unique, appelée "crase", constituée de deux formes simples différentes connues du dictionnaire: καί ἀγαθά → κἀγαθά. Une grammaire constituée de six graphes permet de remplacer les crases par les formes simples correspondantes. La structure de cette grammaire est identique à celle décrite antérieurement (§ 3.2).

La forme κἀγαθά est ainsi remplacée par la séquence {@κἀγαθά,.K} {καί,καί.I+Conj} {ἀγαθά,ἀγαθός.A}. Le premier ensemble entre accolades permet l'affichage dans la concordance de la forme contractée du texte (Figure 5). Le code K (pour "crase", grec κρᾶσις) peut intervenir dans un motif de recherche et permet d'afficher la concordance de toutes les crases d'un corpus. Le deuxième ensemble entre accolades fournit la forme simple du premier élément de la crase, suivie de son lemme (καί) et de sa catégorie morpho-syntaxique (I+Conj = conjonction). La troisième partie fournit la forme simple du second élément de la crase (ἀγαθά) accompagnée du lemme (ἀγαθός) et de sa catégorie morpho-syntaxique (A = adjectif). Cette grammaire identifie et remplace 600 cas de crases.

4-52-2 Διονύσιος μὲν γὰρ ὁ τύραννος ὁ νεώτερος @θοἱμάτιον@ τὸ ἱμάτιον τὸ χρύσεον περιελόμενος
2-24-1 μὴ ἔχοντες καὶ ἄθεοι ἐν τῷ κόσμῳ». {S} Πολλὰ @κἀγαθά@ καὶ ἀγαθὰ γένοιτο τῷ τῶν Σκυθῶν

Figure 5. Extrait de la concordance des crases basée sur la requête <K>

4 L'interface de lemmatisation

Quand ces traitements ont été effectués, l'utilisateur peut activer le module de lemmatisation dont l'interface est constituée de trois parties (Figure 6): (A) cadre des informations lexicales et des boutons de commande; (B) cadre des formes à traiter; (C) cadre des graphes (cfr. §2).

En sélectionnant une forme dans B (ici la forme θεοῦ de la sixième ligne), la phrase s'affiche en caractères gras, les informations lexicales qui lui sont attachées apparaissent en A (θεόω.V; θεός.N+Com) et les graphes de la phrase sont fournis en C1 et C2. Si la forme n'a pas de lemme, une fonction permet de lui en attribuer un. Si la forme a reçu plusieurs propositions de lemme, il suffit de valider la proposition conforme au contexte dans lequel elle

s'actualise. Cette opération peut également être effectuée par suppression manuelle des "boîtes" non conformes dans le graphe placé en C1.

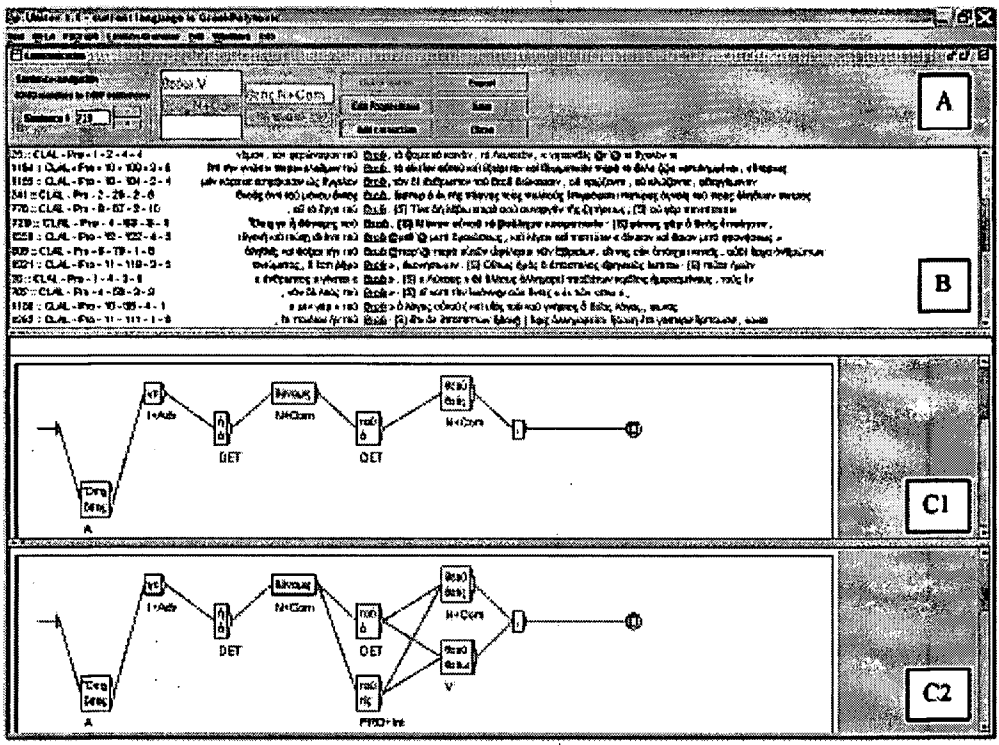

Figure 6. L'interface de lemmatisation

En C2, l'interface produit soit un graphe fixe conservant l'état initial du texte, soit un graphe obtenu après l'application de règles de désambiguïsation. Le travail achevé, une fonction permet d'exporter toutes les données lexicales vers un programme d'édition des outils lexicaux proposés dans les volumes du TPG (concordances, index fréquentiels et inverses, listes lemmes-formes ou formes-lemmes, etc.).

5 Conclusions

Le texte du *Protreptique* (23.062 occurrences) représente 8.827 formes différentes dont 210 cas d'élision et 48 cas de crase. L'outil décrit a étiqueté 8.403 formes différentes (ce qui représente une couverture lexicale de 93,5%). Les formes non reconnues sont principalement des noms propres ou des dérivés de noms propres (l'anthroponyme Εὔμολπος; l'ethnique Συηνῖται, du toponyme Συήνη connu du DAG). Les élisions nouvelles (ἀγαλματ'; ἥμαρτ') et les crases inédites (κἄστι = καί ἔστι; τἄνδον = τὰ ἔνδον) sont traitées manuellement, ce qui permet de mettre à jour les grammaires. Ces dernières, peu à peu enrichies, sont appli-

cables sur d'autres sources. Le traitement du *Protreptique* en est à ce stade. Il sera achevé quand les 2.945 occurrences ambiguës auront été traitées.

Les ressources décrites accompagnent le dictionnaire et en garantissent l'homogénéité et l'économie, rejoignant ainsi un des postulats des concepteurs du DAG: assurer une représentation lexicographique satisfaisante du lexique de la langue. Le dictionnaire ne retient pas, à côté des formes ἄνθρωπος ou ἀγάλματα, des formes du type ἄνθρωπός ou ἀγαλματ΄, représentation contraire à l'habitude des philologues spécialisés dans l'étude du GA. De tels mots sont désormais traités lors de la normalisation du texte.

L'interface de lemmatisation offre un espace de travail conçu pour visualiser les formes à lemmatiser et les ambiguïtés lexicales à réduire. Les données produites sont ensuite récupérables sous d'autres applications. D'une manière plus générale, cette version d'UNITEX constitue déjà un moteur de recherche et un outil d'analyse original dans le domaine du GA. Les travaux futurs intègreront progressivement les informations flexionnelles et les premières grammaires de levée automatique des ambiguïtés lexicales.

Bibliographie

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

Coulie, B. (2003), 'Corpus Christianorum. Thesaurus Patrum Graecorum', in Leemans, J. (ed.) *Corpus Christianorum 1953-2003. Xenium Natalicium. Fifty Years of Scholarly Editing*, Turnhout, pp. 169-172.

Deodati, S. (2005), *Nozioni ed intuizioni linguistiche in Clemente Alessandrino*, tesi di dottorato di ricerca in "Linguistica Storica e Storia Linguistica Italiana", ciclo XVII, Università degli Studi di Roma "La Sapienza" – Libera Università "Maria SS. Assunta" (disponible à l'adresse http://padis.uniroma1.it/getfile.py?recid=301).

Kevers L., Kindt, B. (2004), 'Vers un concordanceur-lemmatiseur en ligne du grec ancien.' *L'Antiquité Classique* 73, pp. 203-213.

Kevers L., Kindt, B. (2005), 'Traitement automatisé de l'ambiguïté lexicale en grec ancien. Première approche par application de grammaires locales.' *Lingvisticae Investigationes* 28, pp. 235-254.

Kindt, B. (2004) 'La lemmatisation des sources patristiques et byzantines au service d'une description lexicale du grec ancien. Les principes de formulation des lemmes du Dictionnaire Automatique Grec (D.A.G.).' *Byzantion* 74, pp. 213-272.

Kindt, B., Yannacopoulou, An. (2006) 'Literary Words Automatic Recognition In a Modern Greek Journalistic *Corpus*', in *Proceedings of the 7th International Conference of Greek Linguistics* (à paraître).

Laporte, É. (2001), 'Reduction of lexical ambiguity.' *Lingvisticae Investigationes* 24, pp. 67-103.

Laporte, É., Monceaux, A. (1998-1999) 'Elimination of lexical ambiguities by grammars: the Elag system.' *Lingvisticae Investigationes* 22, pp. 341-367.

Paumier, S. (2003), *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Thèse de l'Université de Marne-la-Vallée, 2003; à compléter par Paumier, S. (2004).

Paumier, S. (2004), *Unitex 1.2. Manuel d'utilisation*, Université de Marne-la-Vallée.

Verraghenne, C. (2005), *Une affaire de famille? Analyse socio-linguistique de la pièce Iphigénie à Aulis d'Euripide*. Mémoire de Licence, Louvain-la-Neuve.