Bastien KINDT, *Du texte à l'index. L'étiquetage lexical du* De Septem Orbis Spectaculis *de Philon le Paradoxographe: méthode et finalité*, dans *Dialogues d'histoire ancienne*

(Pre-print version published with the courtesy of Professeur Guy Labarre, editor)

How to cite? Please use the traditional bibliographic citation adding the URL of GREgORI project

http://www.uclouvain.be/gregori-project

Softwares, linguistic data and tagged corpus for Ancient **GRE**ek and **ORI**ental languages

Bastien KINDT
Septembre 2015

Du texte à l'index. L'étiquetage lexical du *De Septem Orbis Spectaculis* de Philon le Paradoxographe : méthode et finalité*

1. Le projet GREgORI

Depuis 1991, le *Projet de recherche en lexicologie grecque* (PRLG), mené à l'Institut orientaliste de l'UCL, fournit des outils de traitement automatique pour le grec ancien[1]. Ces outils permettent d'enrichir d'un étiquetage lexical et flexionnel les versions numériques des textes grecs : chaque forme du texte est analysée au niveau lexical (c'est-à-dire identifiée par un lemme univoque accompagné de la mention de sa catégorie morphosyntaxique) et au niveau grammatical (mention des cas, genres, nombres, voix, modes, temps, personnes, etc.). Le but d'une telle opération d'étiquetage est d'obtenir une version lemmatisée et désambiguïsée des textes traités. Ces corpus sont ensuite utilisables à trois fins complémentaires :

1. tirer de ces corpus des index et des concordances lemmatisés, monolingues ou bilingues, offrant une vision exhaustive du vocabulaire d'un texte ou d'un auteur ;

2. constituer un dictionnaire électronique, pierre angulaire des opérations d'étiquetage, en l'occurrence le *Dictionnaire Automatique Grec* (DAG) qui, enrichi au fil des analyses successives, constitue un dictionnaire basé sur des observations réalisées directement sur corpus ;

3. et, enfin, rendre ces corpus consultables et interrogeables par un utilisateur, qu'il soit philologue, linguiste, historien ou autre.

Les textes traités sont principalement, mais non exclusivement, des écrits des Pères de l'Église et des historiens byzantins. À partir de l'année 2013, les outils développés dans le cadre du PRLG sont peu à peu adaptés aux différentes langues de l'Orient chrétien dans lesquelles ces œuvres, surtout celles des Pères, sont susceptibles d'avoir été traduites. À l'heure actuelle, cela concerne en particulier l'arabe, l'arménien et le géorgien. Les développements du PRLG, initialement conçus pour l'analyse du grec, évoluent désormais

* L'auteur tient à remercier ses collègues Marcel Pirard et Hubert Naets qui ont relu ces pages et nous ont fait part de nombreuses suggestions utiles. Cet article est le fruit de travaux menés à l'UCL au sein du projet de recherche 0500230 « Projet de recherche en lexicologie grecque : Réalisation de concordances lemmatisées des auteurs grecs patristiques et byzantins ». Une partie des recherches a également été financées grâce à la subvention Action de recherche concertée (ARC) n°12/17-042 de la Communauté française de Belgique et octroyée par l'Académie universitaire « Louvain ».

[1] Cfr http ://www.uclouvain.be/gregori-project (tous les hyperliens cités sont actifs à l'heure où sont écrites ces lignes). Le lecteur trouvera sur ce site une bibliographie générale du projet.

dans une perspective multilingue. Cette évolution justifie le changement de nom : l'appellation PRLG fait place à « GRE*g*ORI. Softwares, linguistic data and tagged corpus for ancient GREek and ORIental languages ». Cette nouvelle dénomination évoque les langues grecque et orientales, d'une part, et le nom de Grégoire de Nazianze, d'autre part. L'allusion à celui que les Byzantins eux-mêmes appelaient le Théologien n'est pas due au hasard. Historiquement, c'est justement le programme d'édition des *opera omnia* du Nazianzène — programme lui aussi mené à l'Institut orientaliste — qui a donné naissance au PRLG[2].

Si les ressources linguistiques du projet GRE*g*ORI sont conçues par les chercheurs de l'Institut orientaliste, les développements informatiques sont quant à eux pris en charge par le CENTAL, une plate-forme technologique de l'UCL spécialisée dans le traitement automatique des langues[3]. Sans la concertation étroite entre les philologues de l'Institut et les spécialistes du traitement automatique des langues (TAL) du CENTAL, ni le PRLG jadis, ni le projet GRE*g*ORI, aujourd'hui, n'auraient vu le jour. Par ailleurs, certains logiciels issus d'autres laboratoires, dont UNITEX et mkAlign, ont été adoptés et adaptés aux besoins du projet GRE*g*ORI, en concertation avec leur auteur ou les institutions qui les diffusent[4].

L'objectif de cet article est simple : illustrer les traitements réservés aux textes étudiés. Le propos restera volontairement généraliste. De nombreux présupposés techniques ou linguistiques, sans doute importants aux yeux des spécialistes, ne seront pas évoqués dans ces lignes. Le but est de montrer aux utilisateurs potentiels de ces ressources ce qui peut être fait et comment on le fait. Le *De Septem Orbis Spectaculis* (DSOS), tant le texte grec que sa traduction française, servira de fil rouge tout au long de cette présentation. Les traitements effectués sur cette source y sont illustrés pas à pas, depuis la saisie du texte de l'édition, jusqu'aux concordances et aux index bilingues mettant en parallèle le texte grec et sa traduction française[5].

Le DSOS est un texte, tardif et difficilement datable (IV^e^–VI^e^ s.), transmis par deux manuscrits : le Codex Palatinus Graecus 398 (folios 56v-59v ; dernier quart du IX^e^ s.) conservé à Heidelberg et le British Library Add MS 19391 (folios 12v-13v, premier quart du XIV^e^ s.) conservé à Londres, copie du précédent[6]. Son mérite réside dans le fait qu'il est le plus ancien témoin connu réunissant dans un même *opus* la description des sept merveilles du monde : les jardins suspendus de Babylone, les pyramides de Memphis, la statue chryséléphantine de Zeus à Olympie, le Colosse de Rhodes, les remparts de Babylone, le temple d'Artémis à Éphèse et le Mausolée d'Halicarnasse. Certaines descriptions techniques qu'il transmet ont de longue date suscité l'intérêt des chercheurs. Dans les manuscrits, le DSOS est explicitement attribué à Philon de Byzance. Cette information fait songer à l'ingénieur homonyme du III^e^ s. av. J.-C. Mais ce dernier ne peut être l'auteur du DSOS et,

[2] Pour les travaux relatifs à Grégoire de Nazianze, cfr http ://nazianzos.fltr.ucl.ac.be. La concordance des œuvres complètes de cet auteur est parue en deux volumes en 1990 et 1991 (*Thesaurus Gregorii Nazianzeni*).
[3] Sur le CENTAL, cfr http ://www.uclouvain.be/cental.html.
[4] Nous revenons plus loin sur ces deux logiciels, cfr ci-dessous et les notes 15 et 35.
[5] Sur le DSOS, cfr DSOS (Brodersen), p. 14-19 (et la bibliographie, p. 166-167). Pour le texte grec du DSOS, cfr DSOS (Brodersen), p. 20-37 ; pour la traduction française, cfr ROMER —ROMER, *Sept Merveilles*, p. 309-315 (traduction française par D.-A. CANAL).
[6] Cfr DSOS (Brodersen), p. 15-19 ; des copies de ces deux manuscrits sont accessibles en ligne sous les adresses http ://digi.ub.uni-heidelberg.de/diglit/cpgraec398/0116?sid=fc64aacaf345fd0cab2d46857fbc82a2 et http ://www.bl.uk/manuscripts/Viewer.aspx?ref=add_ms_19391_f012v.

par commodité, l'auteur présumé est désigné sous le nom de Philon (de Byzance), dit le Paradoxographe[7].

Le Tableau 1 fournit une évaluation quantitative du lexique du DSOS, tant en grec (GRC) qu'en français (FRA)[8], en terme de mots-occurrences (nombre total des mots du texte), de formes différentes et de lemmes.

DSOS	Mots-occurences	Formes différentes	Lemmes
GRC	1 516	826	589
FRA	2 352	833	686

Tableau 1. Nombre de mots-occurrences, de formes différentes et de lemmes dans le texte grec et la traduction française du DSOS

Le DSOS est un petit texte au regard des œuvres habituellement traitées dans le cadre du projet GRE*g*ORI. À titre de comparaison, le corpus des *opera omnia* de Basile de Césarée, dont la concordance est parue en 2002, compte 707 853 mots-occurrences, 68 867 formes différentes et 14 943 lemmes[9]. Malgré tout, cet *opus minus* convient tout à fait pour illustrer les différentes étapes des traitements réalisés et pour préciser la nature des outils produits à la suite de ces traitements.

2. Préparation du texte pour le traitement informatique

Les textes traités sont tirés d'une édition existante, appelée « édition de référence », ayant fait l'objet, si possible, d'une édition critique. En l'occurrence, il s'agit ici du texte tel qu'il a été édité en 1992 par Kai Brodersen[10]. La Figure 1 en fournit le titre et les premières lignes du texte grec[11].

[7] Sur Philon de Byzance, l'ingénieur du III^e^ s. av. J.-C., cfr M. FOLKERTS, art. *Philon von Byzanz* dans *D.N.P.*, 9, col. 848-849.
[8] Les différentes langues traitées sont désignées par leur code dans la norme ISO 639.2 : GRC pour le grec ancien et FRA pour le français ; cfr http ://www.loc.gov/standards/iso639-2.
[9] *Thesaurus Basilii*, vol. I, p. LI.
[10] Pour les différentes éditions du DSOS, cfr DSOS (Brodersen), p. 166. Le *Thesaurus Graecae Linguae* (dans sa version payante) reprend la même édition du DSOS (TLG 2595.001) ; cfr https ://stephanus.tlg.uci.edu. Nous évoquerons la différence entre les réalisations du projet GRE*g*ORI et celles du TLG dans la conclusion de cet article.
[11] DSOS (Brodersen), p. 20.

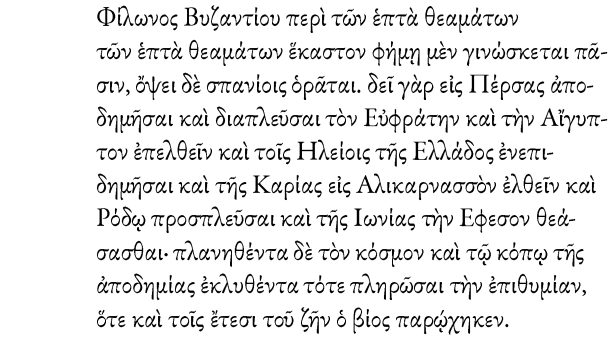
Φίλωνος Βυζαντίου περὶ τῶν ἑπτὰ θεαμάτων
τῶν ἑπτὰ θεαμάτων ἕκαστον φήμῃ μὲν γινώσκεται πᾶ-
σιν, ὄψει δὲ σπανίοις ὁρᾶται. δεῖ γὰρ εἰς Πέρσας ἀπο-
δημῆσαι καὶ διαπλεῦσαι τὸν Εὐφράτην καὶ τὴν Αἴγυπ-
τον ἐπελθεῖν καὶ τοῖς Ηλείοις τῆς Ελλάδος ἐνεπι-
δημῆσαι καὶ τῆς Καρίας εἰς Αλικαρνασσὸν ἐλθεῖν καὶ
Ῥόδῳ προσπλεῦσαι καὶ τῆς Ιωνίας τὴν Εφεσον θεά-
σασθαι· πλανηθέντα δὲ τὸν κόσμον καὶ τῷ κόπῳ τῆς
ἀποδημίας ἐκλυθέντα τότε πληρῶσαι τὴν ἐπιθυμίαν,
ὅτε καὶ τοῖς ἔτεσι τοῦ ζῆν ὁ βίος παρῴχηκεν.

Figure 1. Le titre et les premières lignes du DSOS

Le texte, scanné ou saisi à nouveau frais, reçoit une mise en forme spécifique propre à distinguer, au fil des traitements, les données proprement linguistiques (le texte) et les métadonnées (spécialement les références utiles pour retrouver chaque mot du texte dans l'édition de référence). Ces dernières sont consignées sous des variables numérotées de 0 à 9 et précédées du signe « $ ». D'une manière générale, la première variable ($0=...) mentionne un code ou une abréviation servant à identifier l'auteur ou le corpus traité. La seconde ($1=...) est un identifiant de l'œuvre. Les variables suivantes ($2 à $8), selon l'organisation interne du texte de l'édition, identifient les livres, les chapitres ou les paragraphes constitutifs du texte dans l'édition. La dernière variable ($9=...) indique conventionnellement la numérotation des lignes. La Figure 2 donne un aperçu du texte avec la mention de ces différentes variables.

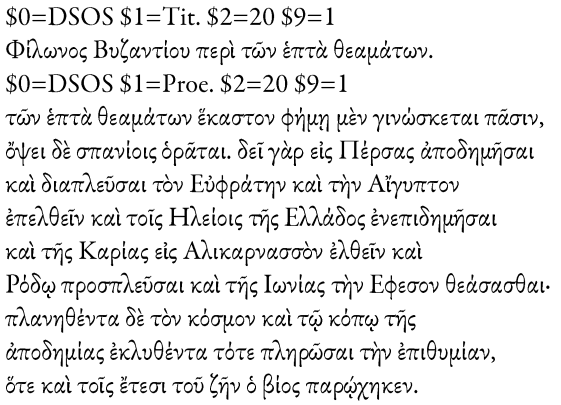
$0=DSOS $1=Tit. $2=20 $9=1
Φίλωνος Βυζαντίου περὶ τῶν ἑπτὰ θεαμάτων.
$0=DSOS $1=Proe. $2=20 $9=1
τῶν ἑπτὰ θεαμάτων ἕκαστον φήμῃ μὲν γινώσκεται πᾶσιν,
ὄψει δὲ σπανίοις ὁρᾶται. δεῖ γὰρ εἰς Πέρσας ἀποδημῆσαι
καὶ διαπλεῦσαι τὸν Εὐφράτην καὶ τὴν Αἴγυπτον
ἐπελθεῖν καὶ τοῖς Ηλείοις τῆς Ελλάδος ἐνεπιδημῆσαι
καὶ τῆς Καρίας εἰς Αλικαρνασσὸν ἐλθεῖν καὶ
Ῥόδῳ προσπλεῦσαι καὶ τῆς Ιωνίας τὴν Εφεσον θεάσασθαι·
πλανηθέντα δὲ τὸν κόσμον καὶ τῷ κόπῳ τῆς
ἀποδημίας ἐκλυθέντα τότε πληρῶσαι τὴν ἐπιθυμίαν,
ὅτε καὶ τοῖς ἔτεσι τοῦ ζῆν ὁ βίος παρῴχηκεν.

Figure 2. Extrait du fichier « DSOS.cental »

Ces variables peuvent être adaptées aux spécificités de chaque texte. Ici, la première variable ($0=DSOS) identifie sans surprise le texte du DSOS. La seconde ($1=Tit. ou $1=*Proe.*) identifie respectivement les parties du texte correspondant au titre (Titulus) et à l'introduction (*Proemium*). Dans les deux manuscrits qui transmettent le texte du DSOS, la description de la sixième merveille (le temple d'Artémis) est incomplète et celle de la septième (le Mausolée d'Halicarnasse) est perdue. Chaque merveille est décrite dans un chapitre du texte, numéroté de 1 à 6 et déclaré par cette variable « $1 ». L'indication « $2=20 » renvoie à la page 20 de l'édition. La dernière variable indique le numéro de la ligne ($9=1). Au besoin, ces variables sont facilement transposables dans un standard de

représentation des textes numérisés conforme aux principes édictés par la Text Encoding Initiative (TEI)[12].

Dans la Figure 2, on remarquera que les quatre formes situées en fin de ligne dans l'édition (ἀποδημῆσαι, Αἴγυπτον, ἐνεπιδημῆσαι et θεάσασθαι sur la Figure 1), et qui sont scindées par un trait d'union pour des raisons évidentes de mise en page (alignement à droite) propre au travail de composition, sont ici reconstituées : la forme ἀπο-|δημῆσαι du texte *édité* devient ἀποδημῆσαι dans le texte *traité*.

Ainsi mise en forme, la version numérisée de l'édition devient un corpus. En pratique, la majeure partie des opérations d'édition des variables et de rattachement des formes scindées est automatisée. Les programmes du CENTAL fournissent un « tableau de bord » permettant de réaliser facilement les multiples opérations utiles. Le corpus — enregistré comme Texte Brut Unicode et marqué de l'extension « .cental » (le nom du fichier est donc « DSOS.cental ») — est maintenant disponible pour traitement[13].

3. Importation du texte dans une base de données

À ce stade, le fichier « DSOS.cental » peut être importé dans une base de données. Tous les éléments du corpus (données et métadonnées) y sont repris sous la forme d'un « fichier vertical » disposé en colonnes. La figure 3 fournit une image volontairement simplifiée de la structure de la base de données[14].

1	2	3	4	5	6	7	8	9
1	DSOS	Tit.	20	1	Φίλωνος	φίλωνος		
2	DSOS	Tit.	20	1	Βυζαντίου	βυζαντίου		
3	DSOS	Tit.	20	1	περὶ	περί		
4	DSOS	Tit.	20	1	τῶν	τῶν		
5	DSOS	Tit.	20	1	ἑπτὰ	ἑπτά		
6	DSOS	Tit.	20	1	θεαμάτων.	θεαμάτων		
7	DSOS	Proe.	20	1	τῶν	τῶν		
8	DSOS	Proe.	20	1	ἑπτὰ	ἑπτά		
9	DSOS	Proe.	20	1	θεαμάτων	θεαμάτων		
10	DSOS	Proe.	20	1	ἕκαστον	ἕκαστον		
11	DSOS	Proe.	20	1	φήμῃ	φήμῃ		
12	DSOS	Proe.	20	1	μὲν	μέν		
13	DSOS	Proe.	20	1	γινώσκεται	γινώσκεται		
14	DSOS	Proe.	20	1	πᾶσιν,	πᾶσιν		
15	DSOS	Proe.	20	2	ὄψει	ὄψει		
16	DSOS	Proe.	20	2	δὲ	δέ		
17	DSOS	Proe.	20	2	σπανίοις	σπανίοις		
18	DSOS	Proe.	20	2	ὁρᾶται.	ὁρᾶται		
19	DSOS	Proe.	20	2	δεῖ	δεῖ		
20	DSOS	Proe.	20	2	γὰρ	γάρ		
21	DSOS	Proe.	20	2	εἰς	εἰς		

[12] Sur la TEI, cfr http ://www.tei-c.org.

[13] Quelques remarques typographiques s'imposent à propos du texte du DSOS dans sa dernière édition : 1°) p. 20, l. 13 : le texte de l'édition a la forme εἶδεν (le texte du TLG — cfr note 10 — a ἴδεν) ; 2) p. 22, l. 2 : le texte de l'édition à la forme δό'ξαν (*sic*) pour δόξαν (le TLG a la forme correcte) ; 3) p. 24, l. 24 : le texte de l'édition et le TLG présentent la forme πυραμίδες qui doit être corrigée en πυραμίδας, comme l'impose la syntaxe de la phrase (c'est d'ailleurs la forme présente dans toutes les éditions antérieures et qui est également bien visible sur les fac-simile cités sous la note 6). Ces corrections ont été prises en compte dans le fichier « DSOS.cental. »

[14] Pratiquement, toutes les données du projet GREgORI sont stockées dans des bases de données relationnelles.

22	DSOS	Proe.	20	2	Πέρσας	πέρσας		
23	DSOS	Proe.	20	2	ἀποδημῆσαι	ἀποδημῆσαι		
24	DSOS	Proe.	20	3	καὶ	καί		
25	DSOS	Proe.	20	3	διαπλεῦσαι	διαπλεῦσαι		
26	DSOS	Proe.	20	3	τὸν	τόν		
27	DSOS	Proe.	20	3	Εὐφράτην	εὐφράτην		
28	DSOS	Proe.	20	3	καὶ	καί		
29	DSOS	Proe.	20	3	τὴν	τήν		
30	DSOS	Proe.	20	3	Αἴγυπτον	αἴγυπτον		
31	DSOS	Proe.	20	4	ἐπελθεῖν	ἐπελθεῖν		
32	DSOS	Proe.	20	4	καὶ	καί		
33	DSOS	Proe.	20	4	τοῖς	τοῖς		
34	DSOS	Proe.	20	4	Ἠλείοις	ἠλείοις		
35	DSOS	Proe.	20	4	τῆς	τῆς		
36	DSOS	Proe.	20	4	Ἑλλάδος	ἑλλάδος		
37	DSOS	Proe.	20	4	ἐνεπιδημῆσαι	ἐνεπιδημῆσαι		
38	DSOS	Proe.	20	5	καὶ	καί		
39	DSOS	Proe.	20	5	τῆς	τῆς		
40	DSOS	Proe.	20	5	Καρίας	καρίας		
41	DSOS	Proe.	20	5	εἰς	εἰς		
42	DSOS	Proe.	20	5	Ἁλικαρνασσὸν	ἁλικαρνασσόν		
43	DSOS	Proe.	20	5	ἐλθεῖν	ἐλθεῖν		
44	DSOS	Proe.	20	5	καὶ	καί		
45	DSOS	Proe.	20	6	Ῥόδῳ	ῥόδῳ		
46	DSOS	Proe.	20	6	προσπλεῦσαι	προσπλεῦσαι		
47	DSOS	Proe.	20	6	καὶ	καί		
48	DSOS	Proe.	20	6	τῆς	τῆς		
49	DSOS	Proe.	20	6	Ἰωνίας	ἰωνίας		
50	DSOS	Proe.	20	6	τὴν	τήν		
51	DSOS	Proe.	20	6	Ἔφεσον	ἔφεσον		
52	DSOS	Proe.	20	6	θεάσασθαι·	θεάσασθαι		
53	DSOS	Proe.	20	7	πλανηθέντα	πλανηθέντα		
54	DSOS	Proe.	20	7	δὲ	δέ		
55	DSOS	Proe.	20	7	τὸν	τόν		
56	DSOS	Proe.	20	7	κόσμον	κόσμον		
57	DSOS	Proe.	20	7	καὶ	καί		
58	DSOS	Proe.	20	7	τῷ	τῷ		
59	DSOS	Proe.	20	7	κόπῳ	κόπῳ		
60	DSOS	Proe.	20	7	τῆς	τῆς		
61	DSOS	Proe.	20	8	ἀποδημίας	ἀποδημίας		
62	DSOS	Proe.	20	8	ἐκλυθέντα	ἐκλυθέντα		
63	DSOS	Proe.	20	8	τότε	τότε		
64	DSOS	Proe.	20	8	πληρῶσαι	πληρῶσαι		
65	DSOS	Proe.	20	8	τὴν	τήν		
66	DSOS	Proe.	20	8	ἐπιθυμίαν,	ἐπιθυμίαν		
67	DSOS	Proe.	20	9	ὅτε	ὅτε		
68	DSOS	Proe.	20	9	καὶ	καί		
69	DSOS	Proe.	20	9	τοῖς	τοῖς		
70	DSOS	Proe.	20	9	ἔτεσι	ἔτεσι		
71	DSOS	Proe.	20	9	τοῦ	τοῦ		
72	DSOS	Proe.	20	9	ζῆν	ζῆν		
73	DSOS	Proe.	20	9	ὁ	ὁ		
74	DSOS	Proe.	20	9	βίος	βίος		
75	DSOS	Proe.	20	9	παρῴχηκεν.	παρῴχηκεν		
…								

Figure 3. Le fichier « DSOS.cental » importé dans la base de données

Les colonnes de la base de données contiennent les informations suivantes:

1. un numéro séquentiel, de 1 à x, identifiant tous les mots du texte, du premier au dernier, ici de Φίλωνος à παρῴχηκεν (pour l'extrait illustré sous les Figures 1 et 2) ;

2. l'identification de l'auteur ou du corpus (donnée équivalant à la variable $0 du fichier « DSOS.cental »), ici « DSOS » ;

3. l'identification de la section du texte (donnée équivalant à la variable $1 du fichier « DSOS.cental »), ici, « Tit. » et « Proe. » et, plus loin dans la base de données, la référence aux six chapitres conservés du texte ;

4. le numéro de la page du texte de l'édition (donnée équivalant à la variable $2 du fichier « DSOS.cental »), ici, l'extrait cité est tiré de la page 20 ;

5. le numéro de la ligne du texte de l'édition (donnée équivalant à la variable $9 du fichier « DSOS.cental »), ici, l'extrait cité s'étend des lignes 1 à 9 ;

6. les formes du texte, classées verticalement, telles qu'elles apparaissent dans l'édition et appelées « formes d'origine » ;

7. les « formes nettoyées », c'est-à-dire les « formes d'origine » dépouillées des majuscules (Φίλωνος devient φίλωνος) et des marques de ponctuation (θεαμάτων. devient θεαμάτων). Les barytons deviennent des oxytons (περὶ devient περί), les accents d'enclise sont supprimés (une fois nettoyée, la forme μεῖζόν qui apparaît devant ἐστιν en DSOS, 4, 30, 23, devient μεῖζον), etc.

Les colonnes 8 et 9 sont réservées aux lemmes et à leur catégorie morphosyntaxique. Ces informations, absentes du fichier « DSOS.cental », n'apparaissent donc pas encore à ce stade. Elles n'y seront versées qu'après l'importation des données de lemmatisation produites à l'aide du logiciel UNITEX.

4. Unitex

Les traitements proprement linguistiques commencent ici. UNITEX est un logiciel conçu, entre autre, pour l'analyse lexicale, grammaticale et syntaxique des corpus[15]. Après avoir chargé un texte, l'utilisateur peut en explorer le contenu en formulant des requêtes. Les réponses à ces requêtes sont affichées sous la forme de concordances[16] ; elles sont donc lisibles au milieu des contextes d'énonciation dans lesquels elles apparaissent. Les recherches peuvent porter sur une forme du texte (par ex. εἶναι ; Figure 4), sur un lemme (βάλλω ; Figure 5), sur une catégorie morphosyntaxique (un anthroponyme ; Figure 6), sur

[15] Le logiciel UNITEX a été conçu par Sébastien Paumier à l'Institut Gaspard Monge de l'Université Paris-Est Marne-la-Vallée, cfr http ://www-igm.univ-mlv.fr/~unitex. UNITEX est distribué gratuitement sous licence LGPL. Voir aussi *UNITEX 3.1beta Manuel d'Utilisation* (accessible en ligne sous l'adresse http ://www-igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf) et, pour une première approche, LAPORTE, *Concordanciers* (accessible en ligne sous l'adresse https ://halshs.archives-ouvertes.fr/halshs-00432571/document). La version standard d'UNITEX est distribuée avec des ressources linguitiques propres aux différentes langues disponibles (l'allemand, l'arabe, le coréen, l'anglais, le finnois, le français, le grec, l'italien, le norvégien, le polonais, le portugais, le russe, le serbe, l'espagnol et le thai). Elle contient donc aussi un échantillon de resssources pour le traitement des langues pratiquées dans le cadre du projet GREgORI, notamment le grec ancien, l'arménien ancien et le géorgien ancien.

[16] Les modes de recherches possibles sont détaillés dans *Unitex 3.1beta Manuel d'Utilisation*, p. 71-80. Pour les concordances, cfr *Unitex 3.1beta Manuel d'Utilisation*, p. 80-85.

des informations flexionnelles et même sur une combinaison de ces différents éléments (par exemple la suite « préposition + déterminant article + nom neutre » sous la Figure 7)[17].

[DSOS-2-26-11] δοκεῖν ὅλου τοῦ κατασκευάσματος μίαν **εἶναι** πέτρας συμφυίαν.
[DSOS-6-36-8] ὥστε τῆς μὲν ἐπιβολῆς τολμηρότερον **εἶναι** τὸν πόνον, τοῦ πόνου δὲ
[DSOS-1-24-1] τήν τε ἀρόσιμον ὑπὲρ κεφαλῆς **εἶναι** τῶν ἐπὶ τοῖς ὑποστύλοις περιπατούντων

Figure 4. Concordance des formes εἶναι

[DSOS-5-34-7] ἐτείχισεν τριακοσίων ἐξήκοντα σταδίων **βαλλομένη** θεμελίωσιν, ὥστε τὴν περίμετρον
[DSOS-6-36-16] λόντων ἐπαπερείδεσθαι πρῶτον μὲν ἔξωθεν **ἐβάλετο** κρηπῖδα δεκάβαθμον διεγείρων
[DSOS-6-36-12] τῶν ὀρυγμάτων καταβιβάσας εἰς ἄπειρον **ἐβάλετο** τὴν κατώρυγα θεμελίωσιν

Figure 5. Concordance des formes du lemme βάλλω

[DSOS-5-34-4] **Σεμίραμις** ἐς βασιλικὴν ἐπλούτησεν ἐπίνοιαν.
[DSOS-Tit.-20-1] **Φίλωνος** Βυζαντίου περὶ τῶν ἑπτὰ θεαμάτων.
[DSOS-3-28-6] Διὸς Κρόνος μὲν ἐν οὐρανῷ, **Φειδίας** δ΄ ἐν Ἤλιδι πατήρ ἐστιν·

Figure 6. Extrait de la concordance des noms propres anthroponymiques

[DSOS-2-26-25] τῆς κορυφῆς στάσις σκοτοῖ τὰς ὄψεις τῶν **εἰς τὰ βάθη** καταθεωρούντων
[DSOS-4-32-26] ὀλίγον ἀναβὰς **ἐπὶ τὸ τέρμα** τῆς ἐλπίδος
[DSOS-5-34-10] περὶ τὴν ἀσφάλειαν τῆς οἰκοδομίας καὶ **περὶ τὰ πλάτη** τῶν μέσων τόπων·

Figure 7. Concordance des séquences introduites par une préposition suivie d'un déterminant article et d'un nom neutre (sans spécification du nombre)

De telles requêtes sont possibles car UNITEX fonctionne avec des dictionnaires applicables au texte, dans notre cas le DAG, le dictionnaire du projet GREgORI. Les dictionnaires utilisés sous UNITEX ont un formalisme très simple appelé DELAF[18]. Les formes y sont enregistrées dans un fichier Texte Brut Unicode et portant l'extension « .dic ». Chaque forme occupe une ligne du fichier. À la suite de la forme, après une virgule, apparaissent le lemme puis, après un point, l'analyse morphosyntaxique du lemme. Enfin, après un double point, sont consignées les informations flexionnelles de la forme. Un extrait du DAG au format DELAF est fourni sous la Figure 8. Ce dictionnaire n'est pas un dictionnaire au sens « traditionnel » du terme, tel un dictionnaire de traduction, mais un outil d'analyse exploité par un programme informatique.

[17] Les catégories morphosyntaxiques et les informations flexionnelles sont représentées à l'aide d'un jeu d'étiquettes décrit sur le site du projet GREgORI, cfr note 1.
[18] Sur les dictionnaires utilisés par UNITEX, cfr *Unitex 3.1beta Manuel d'Utilisation*, p. 41-49. L'abréviation DELA signifie *Dictionnaire Électronique* du *Laboratoire d'Automatique Documentaire et Linguistique* (LADL) avec la finale « -F » pour indiquer qu'il s'agit d'un dictionnaire de formes. Voir la page consacrée au LADL sous l'adresse cfr http ://infolingu.univ-mlv.fr. Sur le format DELAF, cfr COURTOIS, *Dictionnaires électroniques* et SILBERZTEIN, *Dictionnaires électroniques*. Sur le DAG, cfr KINDT, *Principes*.

φειδία,Φειδίας.N+Ant:Vms
φειδίαι,Φειδίας.N+Ant:Dms
φειδίαν,Φειδίας.N+Ant:Ams
φειδίας,Φειδίας.N+Ant:Nms
φειδίου,Φειδίας.N+Ant:Gms
φειδοῖ,φειδόω.V:EÎP3s
φειδοῖ,φειδώ.N+Com:Dfs
φειδοῖς,φειδόω.V:EÎP2s
φείδοισθε,φείδομαι.V:MOP2p
φείδομαι,φείδομαι.V:MÎP1s

Figure 8. Extrait du DAG au format Delaf

UNITEX fonctionne avec des données linguistiques propres à chaque langue traitée. Ces données sont scrupuleusement séparées du programme proprement dit et sont consignées dans des fichiers que l'utilisateur peut ouvrir, lire, corriger et compléter. Les informations linguistiques utilisées sont également « transparentes » : « N+Ant » désigne un nom anthroponymique, « V » un verbe, « O » l'optatif, « s » ou « p » singulier ou pluriel. Toutes les données manipulées ne sont donc ni codées ni encapsulées et restent *disponibles* ; c'est un atout d'UNITEX.

En entrée, UNITEX reçoit un fichier au format XML (illustré sous la Figure 9) généré automatiquement à partir du fichier « DSOS.cental ».

```
<?xml version="1.0" encoding="UTF-16LE"?>
<DOC source="C :\Users\kindt\Documents\00_GREgORI\PROJETS\DSOS\DSOS.cental">
<R0 utxShort="DSOS">
<R1 utxShort="Tit.">
<R2 utxShort="20">
<R9 utxShort="1»><![CDATA[Φίλωνος Βυζαντίου περὶ τῶν ἑπτὰ θεαμάτων.]]></R9></R2></R1>
<R1 utxShort="Proe.">
<R2 utxShort="20">
<R9 utxShort="1»><![CDATA[τῶν ἑπτὰ θεαμάτων ἕκαστον φήμῃ μὲν γινώσκεται πᾶσιν, ]]></R9>
<R9 utxShort="2»><![CDATA[ὄψει δὲ σπανίοις ὁρᾶται. δεῖ γὰρ εἰς Πέρσας ἀποδημῆσαι ]]></R9>
<R9 utxShort="3»><![CDATA[καὶ διαπλεῦσαι τὸν Εὐφράτην καὶ τὴν Αἴγυπτον ]]></R9>
<R9 utxShort="4»><![CDATA[ἐπελθεῖν καὶ τοῖς Ἠλείοις τῆς Ἑλλάδος ἐνεπιδημῆσαι]]></R9>
<R9 utxShort="5»><![CDATA[καὶ τῆς Καρίας εἰς Ἁλικαρνασσὸν ἐλθεῖν καὶ]]></R9>
<R9 utxShort="6»><![CDATA['Ρόδῳ προσπλεῦσαι καὶ τῆς Ἰωνίας τὴν Ἔφεσον θεάσασθαι·]]></R9>
<R9 utxShort="7»><![CDATA[πλανηθέντα δὲ τὸν κόσμον καὶ τῷ κόπῳ τῆς]]></R9>
<R9 utxShort="8»><![CDATA[ἀποδημίας ἐκλυθέντα τότε πληρῶσαι τὴν ἐπιθυμίαν,]]></R9>
<R9 utxShort="9»><![CDATA[ὅτε καὶ τοῖς ἔτεσι τοῦ ζῆν ὁ βίος παρῴχηκεν.]]></R9>
```

Figure 9. Extrait du fichier « DSOS.xml »

Lorsque l'utilisateur demande l'ouverture du texte (en choisissant l'onglet « Text » puis la commande « Open »), Unitex charge le fichier « DSOS.xml » et réalise un prétraitement qui se compose de trois opérations successives[19] :

1. la segmentation du texte en phrase ;

2. le nettoyage de formes particulières par remplacement ;

3. l'application du dictionnaire.

[19] Sur le prétraitement, cfr *Unitex 3.1beta Manuel d'Utilisation*, p. 30-39.

1. La *segmentation du texte en phrase* repose sur une règle de reconnaissance des signes de ponctuation forte de la langue grecque (le point, le point surélevé, le point virgule qui sert de marque interrogative)[20]. Cette règle ajoute au texte un signe de fin de phrase « {S} » à la suite des ponctuations fortes ; le corpus est ainsi segmenté en phrases. Des règles supplémentaires permettent de traiter correctement les cas particuliers, comme les points de suspension, afin qu'UNITEX ne considère pas les trois points successifs comme autant de signes de fin de phrase. Après cette opération, la dernière ligne du DSOS (DSOS, 6, 37, 17) apparaît donc bien sous la forme :

διεγείρων πρὸς βάσιν μετεωροφανὲς καὶ περὶ [...].{S}

et non sous la forme (qui serait aberrante) :

διεγείρων πρὸς βάσιν μετεωροφανὲς καὶ περὶ [.{S}.{S}.{S}].{S}

La Figure 10 fournit les premières lignes du texte du DSOS telles qu'elles apparaissent dans UNITEX après la segmentation du texte en phrases.

Φίλωνος Βυζαντίου περὶ τῶν ἑπτὰ θεαμάτων.{S}
τῶν ἑπτὰ θεαμάτων ἕκαστον φήμῃ μὲν γινώσκεται πᾶσιν,
ὄψει δὲ σπανίοις ὁρᾶται.{S} δεῖ γὰρ εἰς Πέρσας ἀποδημῆσαι
καὶ διαπλεῦσαι τὸν Εὐφράτην καὶ τὴν Αἴγυπτον
ἐπελθεῖν καὶ τοῖς Ηλείοις τῆς Ελλάδος ἐνεπιδημῆσαι
καὶ τῆς Καρίας εἰς Αλικαρνασσὸν ἐλθεῖν καὶ
Ρόδῳ προσπλεῦσαι καὶ τῆς Ιωνίας τὴν Εφεσον θεάσασθαι.{S}
πλανηθέντα δὲ τὸν κόσμον καὶ τῷ κόπῳ τῆς
ἀποδημίας ἐκλυθέντα τότε πληρῶσαι τὴν ἐπιθυμίαν,
ὅτε καὶ τοῖς ἔτεσι τοῦ ζῆν ὁ βίος παρῴχηκεν.{S}

Figure 10. Extrait du fichier « DSOS.xml » tel qu'il apparaît sous Unitex après la segmentation du texte en phrases

Seules les données linguistiques (le texte) s'affichent sur l'écran de l'utilisateur. Mais les métadonnées (les références) sont conservées en mémoire et apparaissent dans les concordances, comme on le voit sous les figures 4-7.

2. Le *nettoyage de formes*[21] concerne les élisions, les crases et les signes critiques. Un jeu de règles remplace la forme élidée « ἐφ΄ » (DSOS, *Proe.*, 20, 18) par la séquence « {@ἐφ΄@,.EL} {ἐπί,ἐπί.I+Prep} ». Ainsi, un utilisateur qui chercherait toute les formes du lemme ἐπί obtiendra également les formes élidées de ce dernier. Les séquences réécrites par des règles sont encadrées d'accolades. Ce point sera précisé dans la suite de l'exposé. Il n'y a pas de crase dans le texte du DSOS. Dans les autres textes, la crase κἀγαθά est remplacée par les formes simples qui la constituent, καὶ et ἀγαθά, et la séquence réécrite contiendra, comme pour les formes élidées, les formes, les lemmes et les catégories morphosyntaxiques correspondantes, bornées par des accolades :

{@κἀγαθά@,.K} {καὶ,καί.I+Part} {ἀγαθά,ἀγαθός.A}.

[20] *Unitex 3.1beta Manuel d'Utilisation*, p. 32-33.
[21] *Unitex 3.1beta Manuel d'Utilisation*, p. 33-34.

Dans l'état actuel des développements linguistiques propres au traitement du grec ancien, ces règles reconnaissent 1 162 formes élidées différentes et 984 crases. L'utilisateur peut atteindre les crases en indiquant dans sa requête le code « K », la crase elle-même (« κἀγαθά »), ou encore une des deux formes simples ou un des deux lemmes constitutifs de cette crase.

Enfin les signes critiques sont effacés, une forme du type « ἄν[θρω]̣ος » est réécrite « ἄνθρωπος ». Nous reviendrons sur ce point.

Ces modifications sont temporaires et limitées à l'environnement d'UNITEX. Elles préparent les textes pour l'opération suivante, l'*application du dictionnaire*. Les outils finaux (concordances, index ou listes) conservent quant à eux les graphies originales, celles présentes dans les textes édités.

3. Lors de l'*application du dictionnaire,* les données de ce dernier sont comparées aux formes du texte[22]. Quand il y a correspondance, les données du dictionnaire (lemme, catégorie morphosyntaxique et analyses flexionnelles) sont attachées à la forme du texte. Les formes élidées et les crases sont consignées dans le dictionnaire, car elles sont prédictibles (elles sont connues et prévisibles). Les formes marquées de signes critiques (telle ἄν[θρω]̣ος) ne le sont pas, car l'apparition de ces signes est aléatoire, dépendant du constat d'un éditeur par rapport à l'histoire critique de son texte. On comprend donc pourquoi l'application du dictionnaire est réalisée sur des formes sans signes critiques ; le dictionnaire connaît la forme ἄνθρωπος, mais n'enregistre en aucun cas les innombrables variations graphiques possibles, telles que ἄν[θρω]̣ος, [ἄν]θρωπος, etc.

Le DAG totalise à ce jour 434 874 *formes* différentes correspondant à 66 952 *lemmes* relevant de toutes les catégories grammaticales, y compris les noms propres et les déterminants numériques, mots souvent négligés par les dictionnaires traditionnels[23]. Il est régulièrement enrichi au fil des analyses nouvelles. Sur un total de 826 formes différentes attestées dans le DSOS (cfr Tableau 1), 92 formes nouvelles (formes pour lesquelles un lemme était déjà enregistré dans le dictionnaire) ont été ajoutées au DAG (soit 11,13% des formes du texte), ainsi que 46 formes inédites (formes relevant d'un paradigme lexical encore absent du DAG, soit 5,56%). Ces résultats montrent que le DAG, dans son état actuel, connaissait déjà 83,3% du vocabulaire du DSOS. Des évaluations portant sur de larges corpus ont montré que le DAG, selon le type de sources abordées, reconnaissait entre 77% et 94% des formes présentes dans un corpus nouveau soumis à l'analyse[24].

C'est à ce stade qu'intervient une des difficultés majeures du traitement automatique des langues : l'ambiguïté. Les dictionnaires au format DELAF, et donc le DAG, sont des dictionnaires dits « à large couverture ». Lors de leur conception, toutes les analyses possibles pour une forme y sont consignées. Ainsi, en français, pour une forme « avions », deux entrées sont prévues, l'une nominale, l'autre verbale :

```
avions,avion.N+Com
avions,avoir.V
```

[22] *Unitex 3.1beta Manuel d'Utilisation*, p. 62-63.
[23] Sur le *D.A.G.*, cfr KINDT, *Principes*, p. 227-234, §9-15 ; KINDT, *Traitement*, p. 115-172.
[24] KINDT, *Traitement*, p. 168-170 ; cfr aussi KINDT — PIRARD, *Couverture*.

Travailler avec de tels dictionnaires permet de ne pas omettre une analyse possible pour un mot et, partant, d'assurer une analyse exhaustive des textes, toutes les analyses possibles étant prévues. Mais la médaille a son revers. Dans le jargon du TAL, l'application du dictionnaire est une opération dite « brutale », réalisée sans analyse préalable du contexte dans lequel apparaissent les mots du texte. Le programme d'exécution mis en œuvre ne fait pas la différence entre le mot « avions » apparaissant dans la phrase « les *avions* restèrent cloués au sol » et « nous *avions* alors des opportunités nombreuses ». Dans de tels cas, le programme se limite à énumérer les différentes analyses possibles, sans prise de décision. À ce stade des traitements, la forme « avions » du texte sera mise ainsi en relation avec les deux analyses possibles consignées dans le dictionnaire. C'est la question de l'ambiguïté. Puisque l'ambiguïté se situe ici au niveau des lemmes, il s'agit plus précisément d'*ambiguïté lexicale*.

L'ambiguïté touche toutes les langues, y compris le grec ancien. Dans les premières lignes du DSOS (cfr Figures 1 et 2), les formes ὄψει, δεῖ et Ῥόδῳ sont ambiguës et correspondent aux entrées suivantes du DAG :

ὄψει,ὄψις (ἡ).N+Com:Dfs/*action de voir, la vue*
ὄψει,ὁράω.V:MÎF2s:EÎF2s/*voir*
δεῖ,δέω (δήσω).V:MÎP2s:EÎP3s:BÎP2s/*lier, attacher*
δεῖ,δέω (δεήσω).V:MÎP2s:EÎP3s:BÎP2s/*manquer, avoir besoin de*
Ῥόδῳ,Ῥόδος.N+Top:Dfs/*Rhodes (l'île)*
Ῥόδῳ,ῥόδον.N+Com:Dns/*la rose*

Dans la majorité des cas, le lecteur humain n'a même pas conscience de ce genre d'ambiguïtés. En lisant le DSOS, un helléniste comprend instinctivement que ὄψει fonctionne comme un nom, que δεῖ signifie « il faut » et que la forme Ῥόδῳ relève ici du lemme toponymique. Mais les programmes informatiques n'ont aucune connaissance intuitive de la langue. À cette étape du traitement, les ambiguïtés sont révélées, mais ne sont pas résolues. Le Tableau 2 fournit la liste des entrées lexicales des 174 mots-occurrences ambigus parmi les 1 516 mots-occurrences qui constituent le texte grec du DSOS.

αἱ,ὁ.DET:Nfp
αἵ,ὅς ἥ ὅ.PRO+Rel:Nfp
αἵ,ὅς ἥ ὅν.PRO+Pos3s:Nfp
αἴγυπτον,Αἴγυπτος (ἡ).N+Top:Afs
αἴγυπτον,Αἴγυπτος (ὁ).N+Ant:Ams
ἄν,ἄν.I+Part
ἄν,ἄνα.I+Intj
ἄν,ἀνά.I+Prep
ἄν,ἐάν.I+Conj
ἄνω,ἄνω (ἀνά).I+Adv
ἄνω,ἄνω (ἀνύω).V:EÎP1s:ESP1s
ἄπειρον,ἄπειρος (πεῖρα).A:Afs:Ams:Ans:Nns:Vns
ἄπειρον,ἄπειρος (πέρας).A:Afs:Ams:Ans:Nns:Vns
ἀπηρεῖσθαι,ἀπαίρω.V:MWR:BWP
ἀπηρεῖσθαι,ἀπερείδω.V:MWR:BWP
ἅπλατον,ἅπλατος (ἅπλετος).A:Ams:Nfs:Nns:Vns:Ans
ἅπλατον,ἅπλατος (πελάζω).A:Ams:Nfs:Nns:Vns:Ans
ἅς,ὅς ἥ ὅ.PRO+Rel:Afp
ἅς,ὅς ἥ ὅν.PRO+Pos3s:Afp
βάρεσι,βᾶρις.N+Com:Dfp
βάρεσι,βάρος.N+Com:Dnp
θεῶν,θεόω.V:EKPNms:EKPVms
κορυφῆς,κορυφεύς.N+Com:Nmp:Vmp
κορυφῆς,κορυφή.N+Com:Gfs
κρήνας,κραίνω.V:EKJNms:EKJVms
κρήνας,κρήνη.N+Com:Afp
μακάριος,μακάριος.A:Nfs:Nms
μακάριος,Μακάριος.N+Ant:Nms
μέλη,μέλη.N+Com:Nfs:Vfs
μέλη,μέλος.N+Com:Anp:Nnp:Vnp
μένει,μένος.N+Com:Dns
μένει,μένω.V:BÎP2s:EÎP3s:MÎP2s
νοῶν,νοέω.V:Nms:Vms
νοῶν,νόος.N+Com:Gmp
ὅ,ὁ.DET:Nms
ὅ,ὅς ἥ ὅ.PRO+Rel:Ans:Nns:Vns
οἱ,ὁ.DET:Nmp
οἳ,ὁ.DET:Nmp
οἳ,ὅς ἥ ὅ.PRO+Rel:Nmp
οἳ,ὅς ἥ ὅν.PRO+Pos3s:Nmp
οἱ,οὗ.PRO+Per3s:Dfs:Dms:Dns
ὀλύμπιος,Ὀλύμπιος (ὁ).N+Ant:Nms
ὀλύμπιος,Ὀλύμπιος α ον.A:Nms
ὅν,ὅς ἥ ὅ.PRO+Rel:Ams
ὅν,ὅς ἥ ὅν.PRO+Pos3s:Ams:Nns
ὀπτῇ,ὀπτός (ἔψω).A:Dfs
πλέον,πλέω.V:EKPAns:EKPNns:EKPVns
πλοῦτος,πλοῦτος (ὁ).N+Com:Nms
πλοῦτος,πλοῦτος (τό).N+Com:Ans:Nns:Vns
ποίαις,ποία.N+Com:Dfp
ποίαις,ποῖος.PRO+Int:Dfp
πόλις, πόλις.N+Com:Nfs
πόλις,Πόλις.N+Ant:Nfs
πρόσωπον, πρόσωπον.N+Com:Ans:Nns:Vns
πρόσωπον,πρόσωπος.N+Com:Ams
ῥιζῶν,ῥίζα.N+Com:Gfp
ῥιζῶν,ῥιζόω.V:EKPNms:EKPVms
ῥοδίοις,Ῥόδιος α ον.A:Dmp:Dnp
ῥοδίοις,ῥόδιος.A:Dfp:Dmp
ῥόδῳ,ῥόδον.N+Com:Dns
ῥόδῳ,Ῥόδος.N+Top:Dfs
σιδήρου,σίδηρος.N+Com:Gfs:Gms
σιδήρου,σιδηρόω.V
σοί,σός.PRO+Pos2s:Dfs:Dms
σοι,σύ.PRO+Per2s:Dfs:Dms:Dns
σοί,σύ.PRO+Per2s:Dfs:Dms:Dns
σταδίων,στάδιον.N+Com:Gnp
σταδίων,στάδιος.A:Gfp:Gmp:Gnp
συμβόλοις,σύμβολον.N+Com:Dnp

βασιλέα,Βασιλεύς.N+Ant:Ams
βασιλέα,βασιλεύς.N+Com:Ams
βίαις,βία (βίαιος).N+Com:Dfp
βίαις,βία (ὁδός).N+Com+eLat:Dfp
βυζαντίου,Βυζάντιον.N+Top:Gfs:Gns
βυζαντίου,Βυζάντιος.A:Gms:Gns
δέ,δέ.I+Part
δέ,ντέ.I+Part+eItal
δεῖ,δέω (δεήσω).V:BÎP2s:EÎP3s:MÎP2s
δεῖ,δέω (δήσω).V:BÎP2s:EÎP3s:MÎP2s
δόξαν,δοκέω.V:EKJAns:EKJNns:EKJVns
δόξαν,δόξα.N+Com:Afs
ἔδει,δέω (δεήσω).V:EÎI3s
ἔδει,δέω (δήσω).V:EÎI3s
ἐπιφανείας,Ἐπιφάνεια (προσ).N+Ant:Gfs
ἐπιφανείας,Ἐπιφάνεια (τοπ).N+Top:Afp:Gfs
ἐπιφανείας,ἐπιφάνεια.N+Com:Afp:Gfs
ἔργον,ἔργον.N+Com:Ans:Nns:Vns
ἔργον,εἴργω.V:EKPAns:EKPNns:EKPVns
ἔργων,εἴργω.V:EKPNms:EKPVms
ἔργων,ἔργον.N+Com:Gnp
ἔχει,ἔχις.N+Com:Afd:Dfs:Nfd:Vfd
ἔχει,ἔχω.V:BÎP2s:EÎP3s:MÎP2s
ἔχει,χέω.V:EÎI3s
ζῆν,ζάω.V:EWP
ζῆν,Ζεύς.N+Ant:Ams
ἦ,ἦ (αἶ).I+Intj
ἦ,ἦ (καί).I+Part
ἠλείοις,Ἠλεῖος (ὁ).N+Ant:Dmp
ἠλείοις,Ἠλεῖος α ον.A:Dmp:Dnp
ἥν,ὅς ἥ ὅ.PRO+Rel:Afs
ἥν,ὅς ἥ ὅν.PRO+Pos3s:Afs
θεοῦ,θεός.N+Com:Gms
θεοῦ,θεόω.V:BYP2s:MYP2s
θεσπέσιον,θεσπέσιος.A:Afs:Ams:Ans:Nns:Vns
θεσπέσιον,Θεσπέσιος.N+Ant:Ams
θεῶν,θεά.N+Com:Gfp
θεῶν,θέα.N+Com:Gfp
θεῶν,θεός.N+Com:Gmp
ὀπτῇ,ὀπτός (ὁράω).A:Dfs
ὀρῶν,ὀρός.N+Com:Gnp
ὀρῶν,ὄρος.N+Com:Gnp
ὅσιον,ὅσιος.A:Ams:Ans:Nns:Vns
ὅσιον,Ὅσιος.N+Ant:Ams
ὅτε, ὅτε.I+Conj
ὅτε,ὅστε.PRO+Rel:Ans:Nns:Vns
ὅτι, ὅτι.I+Conj
ὅτι,ὅστις.PRO+Rel:Ans:Nns:Vns
οὐράνιον,οὐράνιος.A:Ams:Ans:Nns:Vns
οὐράνιον,Οὐράνιος.N+Ant:Ams
οὐρανόν,Οὐρανός.N+Ant:Ams
οὐρανόν,οὐρανός.N+Com:Ams
οὐρανῷ,Οὐρανός.N+Ant:Dms
οὐρανῷ,οὐρανός.N+Com:Dms
ὄψει,ὁράω.V:EÎF2s:MÎF2s
ὄψει,ὄψις (ἡ).N+Com:Dfs
ὄψις,ὄψις (ἡ).N+Com:Nfs
ὄψις,ὄψις (ὁ).N+Com
πείσῃ,πάσχω.V:MÎF2s:MSJ3s
πείσῃ,πείθω.V:MÎF2s:MSJ3s
πελαγία,Πελαγία.N+Ant:Nfs:Vfs
πελαγία,πελάγιος.A:Nfs:Vfs
πέτρα, πέτρα.N+Com:Nfs:Vfs
πέτρα,Πέτρα.N+Top:Nfs:Vfs
πέτρας,πέτρα.N+Com:Afp:Gfs
πέτρας,Πέτρα.N+Top:Gfs
πετρῶν,πέτρα.N+Com:Gfp
πετρῶν,πετρόω.V:EKPNms:EKPVms
πηγάς, πηγάς.N+Com:Nfs
πηγάς,πηγή.N+Com:Afp
πιστόν,πιστός (πείθω).A:Ams:Ans:Nns:Vns
πιστόν,πιστός (πιπίσκω).A:Ams:Ans:Nns:Vns
πλάτη, πλάτη.N+Com:Nfs:Vfs
πλάτη,Πλάτη.N+Top:Nfs:Vfs
πλάτη,πλάτος (τό).N+Com:Anp:Nnp:Vnp
πλέον,πλείων.A:Ans:Nns:Vfs:Vms:Vns
συμβόλοις,σύμβολος.A:Dfp:Dmp:Dnp
σχεδίας,σχεδία.N+Com:Afp:Gfs
σχεδίας,σχέδιος.A:Afp:Gfs
τισι,τις.PRO+Ind:Dfp:Dmp:Dnp
τίσι,τίς.PRO+Int:Dfp:Dmp:Dnp
τοῦ,ὁ.DET:Gms:Gns
τοῦ,τίς.PRO+Int:Gfs:Gms:Gns
τῷ,ὁ.DET:Dms:Dns
τῷ,τίς.PRO+Int:Dfs:Dms:Dns
ὕλην,ὕλη.N+Com:Afs
ὕλην,Ὕλη.N+Prop:Afs
ὑπηρετῶν,ὑπηρετέω.V:Nms:Vms
ὑπηρετῶν,ὑπηρέτης.N+Com:Gmp
φοίνικες,φοῖνιξ (ὁ).N+Com:Nmp:Vmp
φοίνικες,Φοῖνιξ (ὁ).N+Top:Nmp:Vmp
φοίνικες,φοῖνιξ (φοινίκειος).A:Nfp:Nmp:Vfp:Vmp
φοίνικες,Φοῖνιξ ισσα.A:Nmp:Vmp
φοίνικες,φοῖνιξ ισσα.N+Com:Nmp:Vmp
φύσεις,φύσις.N+Com:Afp:Nfp:Vfp
φύσεις,φύω.V:EÎF2s
φυτῶν,φυτόν.N+Com:Gnp
φυτῶν,φυτός.A:Gfp:Gmp:Gnp
χαλκοῦ,χαλκός.N+Com:Gms
χαλκοῦ,χαλκόω.V:BYP2s:MYP2s
χάριτι,χάρις.N+Com:Dfs
χάριτι,Χάρις.N+Prop:Dfs
χοῦν,χέω.V:Nns:Vns
χοῦν,χόος (χέω).N+Com:Ams
χοῦν,χόος (χώννυμι).N+Com:Ams
χοῦν,χόω.V:EWP
ὦ, ὦ.I+Intj
ὦ,εἰμί.V:ESP1s
ὦ,ὦ (ᾱ̃).I+Intj
ὦ,ὦ (τό).N+Lettre:Ans:Dns:Gns:Nns:Vns
ὧν,ὅς ἥ ὅ.PRO+Rel:Gfp:Gmp:Gnp
ὧν,ὅς ἥ ὅν.PRO+Pos3s:Gfp:Gmp:Gnp
ὡς,ὡς (εἰς).I+AdvPr
ὡς,ὡς (ὅς).I+Conj

Tableau 2. Entrées lexicales du DAG correspondant aux mots-occurrences ambigus rencontrés dans le texte du DSOS

Comme nous l'avons déjà illustré ailleurs, les ambiguïtés lexicales peuvent apparaître là où le philologue ne les attend pas. Deux exemples l'illustrent à souhait : les formes τοῦ et δέ.

En grec ancien, la forme τοῦ correspond soit au déterminant article ὁ/*le* (τοῦ,ὁ.DET:Gms:Gns), soit au pronom interrogatif τίς/*qui ?* (τοῦ,τίς.PRO+Int:Gfs:Gms:Gns). Le problème de cette forme est sa très haute fréquence d'apparition dans les textes, autant de formes qu'il s'agira de traiter.

Cas plus surprenant encore pour les philologues classiques, celui de la forme δέ. En grec ancien (classique ou tardif), ce mot est une particule de coordination. Le lemme « δέ.I+Part » fonctionne en corrélation avec la particule μέν, dans l'expression d'un parallélisme ou d'une opposition (μέν ... δέ .../*d'une part ..., d'autre part ...* ou *mais ...*). À partir du XIII^{e} s., les Occidentaux interviennent en Méditerranée orientale et donc dans l'histoire de l'empire byzantin (les Croisés prennent Constantinople en 1204). À partir de cette époque, les textes fournissent des emplois de cette même forme δὲ utilisée non plus comme

particule de coordination, mais comme particule nobiliaire, insérée dans des séquences désignant des noms propres de personnes. Par exemple, Thomas d'Aquin est désigné dans les sources sous la forme Θωμᾶς δὲ Ἄκινος. Dans ce cas, le lemme sera « ντέ.I+Part ». Au moins huit graphies différentes de ce lemme sont attestées : δε (*sic*, sans accent), δέ, νδε (*sic*, sans accent), νδέ, ντε (*sic*, sans accent), ντέ, τε (*sic*, sans accent) et enfin τέ[25]. La particule de coordination δέ se caractérise également par sa haute fréquence d'utilisation dans les textes. Pour obtenir un corpus parfaitement lemmatisé, tous les cas doivent être résolus.

On touche ici à une caractéristique capitale du projet GRE*g*ORI. Le DAG est un dictionnaire « à large couverture » et a été conçu, dès ses origines, pour traiter le vocabulaire de textes dispersés dans une large fourchette chronologique. Les textes abordés sont principalement tirés des littératures patristique et historiographique d'époque byzantine. Parmi les réalisations du PRLG, le plus ancien corpus traité contient les *Opera omnia* de Clément d'Alexandrie (II^{e}-III^{e} s. ap. J.-C.) et le plus récent porte sur l'*Historia Turcobyzantina* de l'historien Doucas (XV^{e} s. ap. J.-C.)[26]. Mais les développements réalisés pour l'analyse de ces ensembles textuels sont également utilisables pour le traitement de textes relevant d'autres genres littéraires, répondants à d'autres niveaux de langue et produits à d'autres époques, y compris les textes classiques ; ce qui était la spécificité du PRLG demeure l'ambition du projet GRE*g*ORI.

À l'issue du prétraitement décrit ci-dessus, le logiciel UNITEX fournit à l'utilisateur un texte lemmatisé (sans être désambiguïsé), mais déjà interrogeable. La nuance entre *lemmatisé et désambiguïsé*, d'une part, et, *lemmatisé sans être désambiguïsé*, d'autre part, est importante. Cela signifie que, à ce stade, une requête demandant les attestations du lemme ῥόδον/*la rose* présentes dans le DSOS affichera encore la réponse fournie par la concordance de la Figure 11.

[DSOS-4-Tit.-30-6]	ὁ ἐν **Ῥόδῳ** κολοσσός. Ῥόδος ἐστὶ πελαγία νῆσος
[DSOS-Proe.-20-6]	καὶ **Ῥόδῳ** προσπλεῦσαι καὶ τῆς Ἰωνίας τὴν Ἔφεσον θεάσασθαι

Figure 11. Concordance du lemme ῥόδον dans le DSOS avant la désambiguïsation lexicale

Or, dans ces phrases, les formes Ῥόδῳ relèvent toutes deux du toponyme (« Rhodes »), et non du nom commun « rose » ; le lemme ῥόδον n'est en fait pas actualisé dans DSOS. UNITEX fournit des outils permettant de réaliser cette opération de désambiguïsation, soit manuellement, soit automatiquement.

5. Les outils de désambiguïsation lexicale sous UNITEX

UNITEX offre à l'utilisateur une représentation graphique du texte, phrase par phrase. La première phrase du DSOS est la suivante :

[25] Cfr *P.L.P.* #7795, où sont énumérées les différentes graphies attestées pour le nom de Thomas d'Aquin.
[26] Cfr *Thesaurus Clementis et Thesaurus Ducae*.

τῶν ἑπτὰ θεαμάτων ἕκαστον φήμῃ μὲν γινώσκεται πᾶσιν, ὄψει δὲ σπανίοις ὁρᾶται./*chacune des sept merveilles est connue de tous, mais elles ne sont réellement vues que par un petit nombre de personnes.*[27]

La Figure 12 en fournit la représentation graphique.

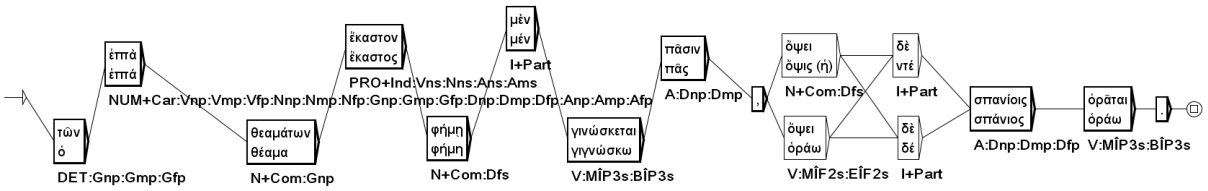

Figure 12. Représentation graphique de la première phrase du DSOS

Dans cette représentation graphique, chaque mot de la phrase, du premier (τῶν) au dernier (ὁρᾶται), est représenté par une boîte. Chaque boîte contient les informations lexicales relatives au mot qu'elle représente : à l'intérieur, la forme du texte et le lemme, puis, sous la boîte, la catégorie morphosyntaxique du lemme et, après deux points, les informations flexionnelles. Ces données proviennent directement du dictionnaire. Les liens entre les boîtes représentent le continuum de la phrase, depuis un état initial symbolisé par la flèche (en amont de la forme τῶν), à l'état final symbolisé par le carré inscrit dans un cercle (après la forme ὁρᾶται et le point final). Quand, sur le même axe vertical, se juxtaposent plus d'une boîte, la représentation graphique de la phrase rend compte d'une ambiguïté lexicale. Nous avons vu que les formes ὄψει et δέ ont différentes analyses lexicales possibles. Pour chacune d'elles, il y a deux boîtes sur le même axe vertical. Le corpus est parfaitement lemmatisé et désambiguïsé quand chaque forme du texte n'est plus représentée que par une et une seule boîte. L'opération de désambiguïsation peut-être réalisée manuellement ou de manière automatisée[28].

La désambiguïsation manuelle est possible. En simplifiant à l'extrême, on peut dire qu'il « suffit » de supprimer les boîtes dont l'analyse ne correspond pas à l'emploi du mot dans la phrase. Si l'opérateur est fiable, le résultat est excellent. Cependant, une telle entreprise, réalisée sur de vastes corpus, devient rapidement chronophage et s'avère donc coûteuse. Le DSOS ne connaît qu'une seule occurrence de la forme ὄψει et quarante-six de la forme δέ. Le corpus de Basile de Césarée (cité ci-dessus) atteste six fois la forme ὄψει, mais connaît 11 521 occurrences de la forme δέ. On objectera, non sans raison, que la particule nobiliaire n'était pas encore entrée dans le lexique de la langue grecque à l'époque des Pères Cappadociens (IV^{e} s.). L'application « brutale » d'un dictionnaire « à large couverture » présente des avantages, mais aussi des inconvénients. Cependant, ôter du dictionnaire les formes de la particule nobiliaire « ντέ.I+Part » reviendrait à fragiliser la cohérence des données et la représentation du lexique de la langue grecque proposée dans le DAG. Fort heureusement, UNITEX propose d'autres approches.

UNITEX est doté de plusieurs outils permettant d'automatiser la désambiguïsation. Une précision s'impose : on parle de désambiguïsation *automatisée* et non *automatique*. L'automatisation de la procédure demeure partielle et, par ailleurs, les résultats doivent

[27] DSOS (Brodersen), *Proe.*, p. 20, l. 1-2 (sauf indication contraire, les traductions sont fournies par l'auteur de cet article).

[28] Nous évitons sciemment d'entrer ici dans les détails des présupposés linguistiques ou techniques sur lesquels se fonde une telle opération.

toujours être soumis à la vérification et à l'approbation d'un philologue à qui revient la charge, *in fine*, de valider ou non les analyses produites par les programmes. Parmi les outils de désambiguïsation, nous en aborderons deux ici :

1. les règles dites « de remplacement » ;

2. le module ELAG.

1. La première de ces deux méthodes repose sur la reconnaissance de séquences de caractères et sur la réécriture de ces séquences par des *règles de remplacement*.

Dans la phrase

καὶ τὸ μὲν ὕψος ἐστὶ **τοῦ τείχους** πλέον ἢ πεντήκοντα πήχεων/*et la hauteur du mur est supérieure à cinquante coudées*[29]

la forme τοῦ attestée devant τείχους ne peut fonctionner que comme déterminant article, et non comme pronom. Il est possible d'écrire une règle selon laquelle, pour la forme τοῦ précédant τείχους, seul le lemme « ὁ.DET:Gns » doit être retenu. Une telle règle est visualisable sous une représentation graphique analogue à celle qui permet de visualiser les phrases du texte. Ainsi, le traitement de la séquence τοῦ τείχους sera réalisé grâce à la règle illustrée sous la Figure 13.

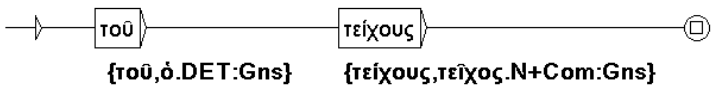

Figure 13. Représentation graphique de la règle de remplacement traitant la séquence τοῦ τείχους

Cette règle se lit de gauche à droite : quand les séquences de caractères τοῦ et τείχους se rencontrent dans un texte parcouru par cette règle, ces deux séquences sont remplacées par les informations consignées entre accolades : « τοῦ» est réécrit « {τοῦ,ὁ.DE :Gns} » et « τείχους » « {τείχους,τεῖχος.N+Com:Gns} ».

La règle intervient sur la séquence de caractères avant même l'application du dictionnaire. En effet, les règles de ce type sont appliquées au texte durant la phase de prétraitement, pendant l'opération de *Nettoyage de formes particulières par remplacement*. À la suite de cette opération, la phrase fournie par UNITEX sera alors

καὶ τὸ μὲν ὕψος ἐστὶ **{τοῦ,ὁ.DET :Gns} {τείχους,τεῖχος.N+Com :Gns}** πλέον ἢ πεντήκοντα πήχεων.

et, rappelons-le, la présence des accolades rend ces formes insensibles à l'application du dictionnaire, opération postérieure à celle du *Nettoyage de formes particulières par remplacement*. La forme τοῦ, en cette position, ne sera donc plus pourvue d'un double étiquetage.

[29] DSOS (Brodersen), chap. 5, p. 34, l. 12.

L'avantage de cette méthode est de fournir des résultats très sûrs. Le désavantage, est que la règle qui traite la séquence τοῦ τείχους ne traite que cette seule séquence. Dans la phrase

τὸ γὰρ χώνευμα **τοῦ κατασκευάσματος** ἐγένετο χαλκούργημα τοῦ κόσμου./*en effet, l'œuvre fondue de la structure devint une oeuvre d'art du monde.*[30]

le syntagme τοῦ κατασκευάσματος, qui répond strictement aux mêmes critères grammaticaux que la séquence τοῦ τείχους, ne sera pas pris en compte, puisque la séquence de caractères diffère. En définitive, il faudrait décrire toutes les séquences possibles constituées de la forme τοῦ suivie directement d'un nom pour résoudre toutes les formes τοῦ ambiguës en cette position : solution irréaliste.

2. La seconde méthode utilise ELAG, un module d'UNITEX. ELAG est l'acronyme de « Elimination of Lexical Ambiguities by Grammars »[31]. Son fonctionnement repose sur les informations lexicales, catégorielles et flexionnelles des mots constituant un syntagme dont l'un des éléments, au moins, doit être désambiguïsé. Puisque cette méthode fait appel aux lemmes, aux catégories morphosyntaxiques et aux analyses flexionnelles, elle ne peut être mise en œuvre qu'après l'application du dictionnaire, contrairement aux règles de remplacement illustrées ci-dessus. Comme son nom l'indique, ELAG travaille avec des règles appelées « grammaires » : nous parlerons désormais de « grammaire ELAG »[32]. Ces grammaires n'interviennent plus sur des séquences de caractères, mais sur les données visibles par l'utilisateur sur la représentation graphique des phrases du texte, comme celle fournie sous la Figure 12.

Revenons aux syntagmes τοῦ τείχους et τοῦ κατασκευάσματος. Les Figures 14 et 15 donnent la représentation graphique de ces syntagmes tels qu'ils apparaîtraient dans la représentation graphique des phrases dans lesquels ils interviennent.

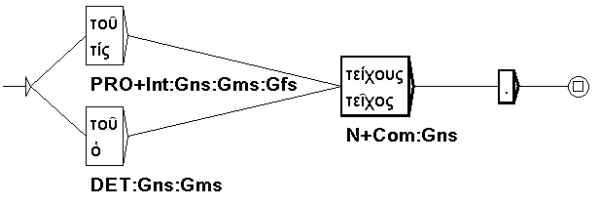

Figure 14. Représentation graphique de la séquence τοῦ τείχους

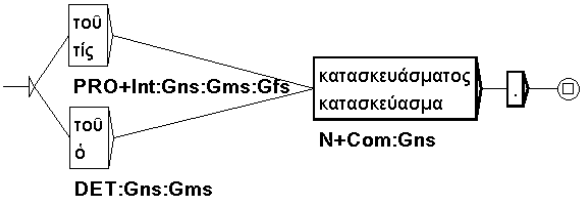

Figure 15. Représentation graphique de la séquence τοῦ κατασκευάσματος

La grammaire ELAG permettant de traiter l'ambiguïté de la forme τοῦ présente dans ces syntagmes dira ceci : SI une forme τοῦ (soit déterminant article, soit pronom interrogatif) précède un nom commun au génitif neutre singulier, ALORS seule l'analyse correspondant au déterminant article doit être conservée dans la représentation graphique de la phrase. La Figure 16 présente une telle grammaire.

[30] DSOS (Brodersen), chap. 4, p. 30, l. 15.
[31] *Unitex 3.1beta Manuel d'Utilisation*, p. 165-179.
[32] Pour une description plus précise de ces grammaires ELAG, cfr LAPORTE — MONCEAUX, *Elag* ; LAPORTE, *Réduction.*

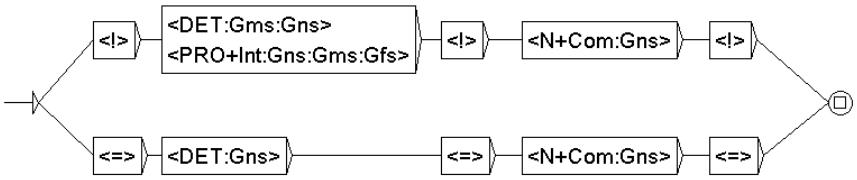

Figure 16. Représentation graphique de la grammaire ELAG *traitant le syntagme τοῦ τείχους*

Visuellement, cette grammaire ELAG propose deux chemins parallèles : un chemin supérieur borné par les signes « <!> » et qui correspond à la condition énoncée au paragraphe précédent, et un chemin inférieur, borné par les signes « <=> », qui décrit l'opération réalisée si la condition se vérifie ; si c'est le cas, seule l'analyse de τοῦ comme déterminant article est conservée et, dans le cas contraire, rien ne change et l'ambiguïté sera résolue ultérieurement, lors d'un contrôle manuel.

Un menu de l'interface ELAG permet à l'utilisateur de demander la compilation de cette grammaire. Pour exprimer les choses simplement, le contenu de la grammaire devient alors une « application informatique » qui réalise l'opération de désambiguïsation souhaitée. Les Figures 17 et 18 en illustrent le résultat sur les syntagmes τοῦ τείχους et τοῦ κατασκευάσματος.

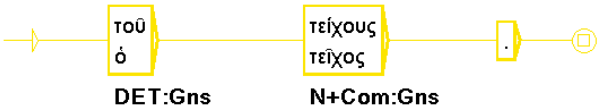

Figure 17. Représentation graphique du syntagme τοῦ τείχους désambiguïsé par la grammaire ELAG *de la Figure 16*

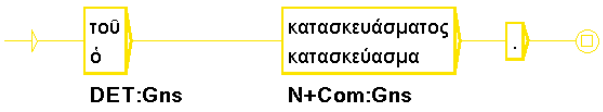

Figure 18. Représentation graphique du syntagme τοῦ κατασκευάσματος désambiguïsé par la grammaire ELAG *de la Figure 16*

Trois remarques s'imposent ici. D'abord, rappelons que les grammaires ELAG n'interviennent pas sur des séquences de caractères, mais sur le texte étiqueté. Cette intervention est donc directement visible sur la représentation graphique de la phrase. Dans ce cas précis, elle ne conserve qu'une seule boîte pour la forme τοῦ — celle correspondant à l'analyse attendue dans un pareil contexte par le concepteur de la grammaire ELAG — et elle élimine le chemin correspondant à l'analyse jugée inadéquate. Ensuite, la grammaire apparie en cas, genre et nombre le déterminant article et le nom. Enfin, la grammaire ELAG travaille avec les étiquettes grammaticales et ne mentionne explicitement aucun lemme. Par conséquent, la même règle vaut tout autant pour le syntagme τοῦ τείχους que pour τοῦ κατασκευάσματος. La généricité des grammaires ELAG est de loin supérieure à celle des règles de remplacement illustrées ci-dessus[33].

[33] UNITEX permet d'utiliser de différentes manières les règles de remplacement. C'est par souci de clarté, ici encore, que nous n'abordons pas la question des règles avec variables (*Unitex 3.1beta Manuel d'Utilisation*, p. 140-145) et celle des règles de remplacement appliquées en cascade sur un texte (*Unitex 3.1beta Manuel d'Utilisation*, p. 245-255).

Dans la réalité des textes, d'autres mots peuvent s'insérer entre le déterminant article et le nom avec lequel il fonctionne : une particule, un adverbe, un adjectif, un participe, etc., ou même une combinaison de ces mots. Pour traiter de telles séquences, il « suffit » de compléter la grammaire ELAG de la Figure 16 de manière à ce que cette dernière puisse tenir compte des autres combinaisons possibles. La grammaire de la Figure 19 remplit ce rôle. Elle prévoit la présence facultative (c'est le sens de l'indication « <E> » dans les boîtes) d'une particule (<I+Part>) à la suite du déterminant article et celle, ici encore facultative, d'un adjectif (<A>) épithète du nom, ce dernier mot clôturant le syntagme. De plus, l'analyse tient compte désormais des noms masculins et assume l'accord en cas, genre et nombre du déterminant article et de l'adjectif avec ces noms, tant au masculin qu'au neutre. L'amélioration des grammaires peut-être réalisée par l'utilisateur puisque, nous le rappelons, ce dernier peut corriger et compléter toutes les ressources linguistiques mises en œuvre sous UNITEX.

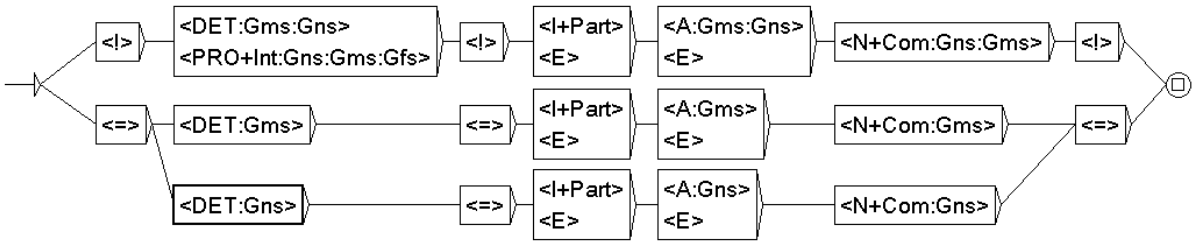

Figure 19. Version complétée de la grammaire ELAG de la Figure 16

Comme l'illustre les figures 20-23, une telle grammaire traite l'ambiguïté lexicale de la forme τοῦ placée en tête de séquence dans différentes expressions :

- τοῦ κατασκευάσματος ;
- τοῦ κολοσσοῦ ;
- τοῦ μεγάλου κατασκευάσματος ;
- τοῦ μεγάλου κολοσσοῦ ;
- τοῦ δὲ κατασκευάσματος ;
- τοῦ δὲ κολοσσοῦ ;
- τοῦ δὲ μεγάλου κατασκευάσματος ;
- ou, enfin, τοῦ δὲ μεγάλου κολοσσοῦ.

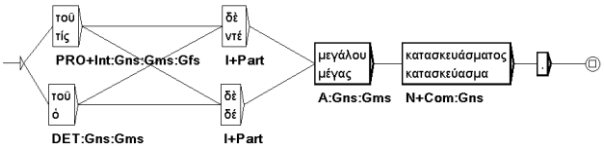

Figure 20. Représentation graphique du syntagme τοῦ δὲ μεγάλου κατασκευασματος

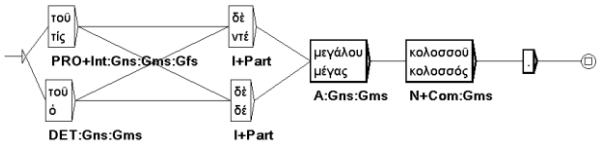

Figure 21. Représentation graphique du syntagme τοῦ δὲ μεγάλου κολοσσοῦ

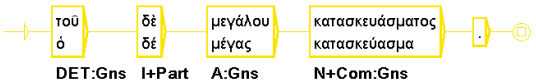

Figure 22. Représentation graphique du syntagme τοῦ δὲ μεγάλου κατασκευασματος désambiguïsé par la grammaire ELAG de la Figure 14

Figure 23. Représentation graphique du syntagme τοῦ δὲ μεγάλου κολοσσοῦ désambiguïsé par la grammaire ELAG de la Figure 14

Quant à la forme δέ, particule de coordination ou particule nobiliaire, elle est traitée par la grammaire ELAG représentée sous la Figure 24. Cette grammaire impose l'interprétation de la forme δέ comme particule de coordination quand elle est précédée d'un mot qui n'est pas un nom propre ou un anthroponyme. Dans les cas contraires, aucune décision n'est prise.

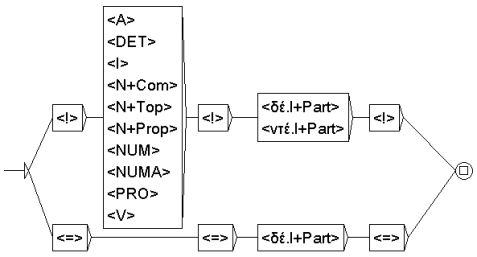

Figure 24. Représentation graphique de la grammaire ELAG traitant l'ambiguïté de la forme δέ

UNITEX permet d'utiliser plusieurs grammaires ELAG en même temps. Dans l'état actuel des développements, un jeu de cent vingt-sept grammaires ELAG assure le traitement de l'ambiguïté lexicale en grec ancien. Ces traitements prévoient la désambiguïsation lexicale de mots relevant de toutes les catégories morphosyntaxiques.

La désambiguïsation lexicale sous UNITEX est réalisée à l'aide de règles de remplacement et à l'aide de grammaires ELAG. Ces deux types de ressources sont mises en œuvre sur les textes. Elles ne suffisent pas pour résoudre la totalité des ambiguïtés d'un texte, mais elles garantissent des traitements fiables et allègent sensiblement les opérations manuelles jadis indispensables pour réaliser le traitement complet d'un corpus. Désormais, seules les ambiguïtés restantes sont traitées à la main.

6. Importation des données de lemmatisation dans la base de données

Une fois qu'un corpus est complètement traité, les données lemmatisées et désambiguïsées doivent être jointes à celles de la base de données illustrée sous la Figure 3. UNITEX dispose d'une fonction d'exportation prévue à cet effet. Des programmes spécifiques réalisés par les informaticiens du Cental récupèrent alors ces informations et les injectent dans la base de données. La Figure 25 affiche la base de données de la Figure 3 enrichie des informations de lemmatisation. Les formes ὄψει (ligne 15 du tableau), δέ (ligne 16) et Ῥόδῳ (ligne 45) sont accompagnées des lemmes qui leur conviennent *in textu*.

1	2	3	4	5	6	7	8	9
1	DSOS	Tit.	20	1	Φίλωνος	φίλωνος	Φίλων	N+Ant
2	DSOS	Tit.	20	1	Βυζαντίου	βυζαντίου	Βυζάντιος	A
3	DSOS	Tit.	20	1	περὶ	περί	περί	I+Prep
4	DSOS	Tit.	20	1	τῶν	τῶν	ὁ	DET
5	DSOS	Tit.	20	1	ἑπτὰ	ἑπτά	ἑπτά	NUM+Car
6	DSOS	Tit.	20	1	θεαμάτων.	θεαμάτων	θέαμα	N+Com
7	DSOS	Proe.	20	1	τῶν	τῶν	ὁ	DET
8	DSOS	Proe.	20	1	ἑπτὰ	ἑπτά	ἑπτά	NUM+Car
9	DSOS	Proe.	20	1	θεαμάτων	θεαμάτων	θέαμα	N+Com
10	DSOS	Proe.	20	1	ἕκαστον	ἕκαστον	ἕκαστος	PRO+Ind
11	DSOS	Proe.	20	1	φήμῃ	φήμῃ	φήμη	N+Com
12	DSOS	Proe.	20	1	μὲν	μέν	μέν	I+Part
13	DSOS	Proe.	20	1	γινώσκεται	γινώσκεται	γιγνώσκω	V
14	DSOS	Proe.	20	1	πᾶσιν,	πᾶσιν	πᾶς	A
15	DSOS	Proe.	20	2	ὄψει	ὄψει	ὄψις (ἡ)	N+Com
16	DSOS	Proe.	20	2	δὲ	δέ	δέ	I+Part
17	DSOS	Proe.	20	2	σπανίοις	σπανίοις	σπάνιος	A
18	DSOS	Proe.	20	2	ὁρᾶται.	ὁρᾶται	ὁράω	V
19	DSOS	Proe.	20	2	δεῖ	δεῖ	δέω (δεήσω)	V
20	DSOS	Proe.	20	2	γὰρ	γάρ	γάρ	I+Part
21	DSOS	Proe.	20	2	εἰς	εἰς	εἰς	I+Prep
22	DSOS	Proe.	20	2	Πέρσας	πέρσας	Πέρσης (Περσίς)	A
23	DSOS	Proe.	20	2	ἀποδημῆσαι	ἀποδημῆσαι	ἀποδημέω	V
24	DSOS	Proe.	20	3	καὶ	καί	καί	I+Part
25	DSOS	Proe.	20	3	διαπλεῦσαι	διαπλεῦσαι	διαπλέω	V
26	DSOS	Proe.	20	3	τὸν	τόν	ὁ	DET
27	DSOS	Proe.	20	3	Εὐφράτην	εὐφράτην	Εὐφράτης	N+Top
28	DSOS	Proe.	20	3	καὶ	καί	καί	I+Part
29	DSOS	Proe.	20	3	τὴν	τήν	ὁ	DET
30	DSOS	Proe.	20	3	Αἴγυπτον	αἴγυπτον	Αἴγυπτος (ἡ)	N+Top
31	DSOS	Proe.	20	4	ἐπελθεῖν	ἐπελθεῖν	ἐπέρχομαι	V
32	DSOS	Proe.	20	4	καὶ	καί	καί	I+Part
33	DSOS	Proe.	20	4	τοῖς	τοῖς	ὁ	DET
34	DSOS	Proe.	20	4	Ἠλείοις	ἠλείοις	Ἦλεῖος (ὁ)	N+Ant
35	DSOS	Proe.	20	4	τῆς	τῆς	ὁ	DET
36	DSOS	Proe.	20	4	Ἑλλάδος	ἑλλάδος	Ἑλλάς	N+Top
37	DSOS	Proe.	20	4	ἐνεπιδημῆσαι	ἐνεπιδημῆσαι	ἐνεπιδημέω	V
38	DSOS	Proe.	20	5	καὶ	καί	καί	I+Part
39	DSOS	Proe.	20	5	τῆς	τῆς	ὁ	DET
40	DSOS	Proe.	20	5	Καρίας	καρίας	Καρία	N+Top
41	DSOS	Proe.	20	5	εἰς	εἰς	εἰς	I+Prep
42	DSOS	Proe.	20	5	Ἁλικαρνασσὸν	ἁλικαρνασσόν	Ἁλικαρνασσός	N+Top
43	DSOS	Proe.	20	5	ἐλθεῖν	ἐλθεῖν	ἔρχομαι	V
44	DSOS	Proe.	20	5	καὶ	καί	καί	I+Part
45	DSOS	Proe.	20	6	Ῥόδῳ	ῥόδῳ	Ῥόδος	N+Top
46	DSOS	Proe.	20	6	προσπλεῦσαι	προσπλεῦσαι	προσπλέω	V
47	DSOS	Proe.	20	6	καὶ	καί	καί	I+Part
48	DSOS	Proe.	20	6	τῆς	τῆς	ὁ	DET
49	DSOS	Proe.	20	6	Ἰωνίας	ἰωνίας	Ἰωνία	N+Top
50	DSOS	Proe.	20	6	τὴν	τήν	ὁ	DET
51	DSOS	Proe.	20	6	Ἔφεσον	ἔφεσον	Ἔφεσος	N+Top
52	DSOS	Proe.	20	6	θεάσασθαι·	θεάσασθαι	θεάομαι (–άω)	V
53	DSOS	Proe.	20	7	πλανηθέντα	πλανηθέντα	πλανάω	V
54	DSOS	Proe.	20	7	δὲ	δέ	δέ	I+Part
55	DSOS	Proe.	20	7	τὸν	τόν	ὁ	DET
56	DSOS	Proe.	20	7	κόσμον	κόσμον	κόσμος	N+Com
57	DSOS	Proe.	20	7	καὶ	καί	καί	I+Part
58	DSOS	Proe.	20	7	τῷ	τῷ	ὁ	DET
59	DSOS	Proe.	20	7	κόπῳ	κόπῳ	κόπος	N+Com
60	DSOS	Proe.	20	7	τῆς	τῆς	ὁ	DET
61	DSOS	Proe.	20	8	ἀποδημίας	ἀποδημίας	ἀποδημία	N+Com
62	DSOS	Proe.	20	8	ἐκλυθέντα	ἐκλυθέντα	ἐκλύω	V
63	DSOS	Proe.	20	8	τότε	τότε	τότε	I+Adv

64	DSOS	Proe.	20	8	πληρῶσαι	πληρῶσαι	πληρόω	V
65	DSOS	Proe.	20	8	τὴν	τήν	ὁ	DET
66	DSOS	Proe.	20	8	ἐπιθυμίαν,	ἐπιθυμίαν	ἐπιθυμία	N+Com
67	DSOS	Proe.	20	9	ὅτε	ὅτε	ὅτε	I+Conj
68	DSOS	Proe.	20	9	καὶ	καί	καί	I+Part
69	DSOS	Proe.	20	9	τοῖς	τοῖς	ὁ	DET
70	DSOS	Proe.	20	9	ἔτεσι	ἔτεσι	ἔτος	N+Com
71	DSOS	Proe.	20	9	τοῦ	τοῦ	ὁ	DET
72	DSOS	Proe.	20	9	ζῆν	ζῆν	ζάω	V
73	DSOS	Proe.	20	9	ὁ	ὁ	ὁ	DET
74	DSOS	Proe.	20	9	βίος	βίος	βίος	N+Com
75	DSOS	Proe.	20	9	παρῴχηκεν.	παρῴχηκεν	παροίχομαι	V
…								

Figure 25. La base de données de la Figure 3 complétée après lemmatisation et désambiguïsation du texte

Les concordances et les index, sous leurs différentes formes, peuvent alors être produits. D'autres programmes fournis par le CENTAL permettent d'extraire de ces bases de données les informations pertinentes (formes, lemmes, éléments d'analyse, références, etc.) et de les injecter dans des formulaires *ad hoc* pour générer les outils lexicologiques dont nous parlons ci-dessous.

7. Concordances et outils lexicologiques monolingues ou bilingues

Les développements du projet GRE*g*ORI sont à même de produire un grand nombre d'outils lexicologiques, comprenant des concordances, des index ou des listes. Trois outils sont téléchargeables librement sur le site du projet GRE*g*ORI[34] :

1. la concordance lemmatisée du texte complet du DSOS ;

2. la liste alphabétique des lemmes et des formes du texte complet du DSOS ;

3. la concordance bilingue (grec ancien – français) du texte complet du DSOS ;

1. *La concordance lemmatisée* du texte complet du DSOS (*Concordantia Lemmatum et Formarum*) : la concordance est un « super index » énumérant toutes les formes du texte classées sous le lemme qui leur correspond. Lemmes et formes sont accompagnés d'une indication de leur fréquence. Les formes sont munies de leurs références et encadrées par leur contexte d'apparition, en amont (contexte à gauche) comme en aval (contexte à droite). La concordance se présente selon l'ordre alphabétique des lemmes. Les formes sont elles aussi classées dans l'ordre alphabétique. Les formes identiques sont classées dans l'ordre alphabétique des lettres du ou des mots qui sui(ven)t, ce qui met en évidence les contextes parallèles. Ce mode de classement choisi peut être modifié selon les attentes de l'utilisateur. Les crases (quand il y en a) sont classées sous les deux lemmes qui les constituent : la crase κἀγαθά apparaît tant sous le lemme καί que sous le lemme ἀγαθός. Les lemmes des déterminants numériques sont précédés de l'indication « N– » et placés en fin de concordance.

[34] Cfr note 1.

2. *La liste alphabétique des lemmes et des formes* (*Enumeratio lemmatum et formarum*) : cette liste est un index sans référence, elle constitue l'inventaire exhaustif du vocabulaire d'un texte ou d'un corpus. Elle fournit la liste alphabétique des lemmes et des formes avec une indication de leur fréquence. Toutes les formes sont orthographiées en minuscules. Les formes sans accent (quand il y en a) proviennent des titres orthographiés en majuscules non accentuées dans les éditions. Chaque lemme est également accompagné de l'indication de sa catégorie morphosyntaxique. Les crases (quand il y en a) sont classées sous les lemmes qui les constituent, comme dans la concordance lemmatisée. De même, les lemmes des déterminants numériques sont précédés de l'indication « N– » et placés en fin de liste.

3. *La concordance bilingue* du texte complet du DSOS. La concordance bilingue propose l'alignement du texte grec (le texte *source*) sur sa traduction française (le texte *cible*). Les concordances bilingues sont créées en deux étapes. Dans un premier temps, les concordances des textes de la langue source et de la langue cible sont réalisées séparément et suivent les différentes étapes des procédures illustrées ci-dessus. Ensuite, les données lemmatisées des deux textes sont introduites dans un logiciel d'alignement bi-textuel, en l'occurrence mkAlign[35]. Ce logiciel permet d'apparier les mots ou expressions d'un texte source sur les mots et expressions correspondants d'un texte cible. Le texte cible peut être une traduction, mais aussi une adaptation ou une contraction d'un texte original (il est en effet possible de comparer deux états différents d'un même texte grec). Les figures 26 et 27 présentent l'alignement grec-français de deux extraits du DSOS tirés du chapitre décrivant la statue chryséléphantine de Zeus à Olympie :

Διὸς Κρόνος μὲν ἐν οὐρανῷ, Φειδίας δ´ ἐν Ἤλιδι πατήρ ἐστιν·/*Kronos est le père de Zeus dans le Ciel, mais Phidias est le père du Zeus de l' Élide.*[36]

μακάριος ὁ καὶ θεασάμενος τὸν βασιλέα τοῦ κόσμου μόνος καὶ δεῖξαι δυνηθεὶς ἄλλοις τὸν κεραυνοῦχον/*Béni soit ce mortel sur la terre, qui a vu le souverain et qui a eu la capacité de montrer le Tonnant aux autres hommes!*[37]

[35] Le logiciel mkAlign a été conçu par Serge Fleury de l'Université Sorbonne Nouvelle Paris 3 ; cfr http ://www.tal.univ-paris3.fr/mkAlign. Tout comme UNITEX, mkAlign est distribué librement pour les chercheurs. UNITEX possède également une fonction d'alignement de texte (cfr *Unitex 3.1beta Manuel d'Utilisation*, p. 207-213) qui n'a pas été retenue dans le cadre du PRLG et du projet GREgORI.
[36] DSOS (Brodersen), chap. 3, p. 28, l. 1-2 (Traduction de D.-A. Canal dans ROMER — ROMER, *Sept Merveilles*, p. 311).
[37] DSOS (Brodersen), chap. 3, p. 28, l. 3-52 (Traduction de D.-A. Canal dans ROMER — ROMER, *Sept Merveilles*, p. 312).

	de
Διὸς	Zeus
Κρόνος	Kronos
μὲν	
ἐν	dans
	le
οὐρανῷ,	Ciel,
Φειδίας	Phidias
δ΄	mais
	le
ἐν	de
	l'
Ἤλιδι	Élide.
	le
πατήρ	père
ἐστιν·	est

Figure 26. Exemple d'alignement de l'extrait Διὸς Κρόνος μὲν...

μακάριος	Béni
	soit
ὁ	ce mortel sur la terre
καὶ	
θεασάμενος	qui a vu
τὸν	le
βασιλέα	souverain
τοῦ	
κόσμου	
μόνος	
καὶ	et
δεῖξαι	de montrer
δυνηθεὶς	qui a eu la capacité
	aux
ἄλλοις	autres hommes!
τὸν	le
κεραυνοῦχον.	Tonnant

Figure 27. Exemple d'alignement de l'extrait μακάριος ὁ καὶ...

Dans certains cas, il est facile d'apparier mot à mot le texte source et le texte cible. C'est le cas pour la première phrase (Figure 26). Les seules différences s'observent à deux niveaux :

1) celui des particules du grec (μέν, δ΄) qui n'existent pas en français. L'opposition qu'elles expriment est rendue par le seul mot « mais » ;

2) celui des déterminants articles (trois fois « le », une fois « l' »), dont les emplois sont fort différents en grec et en français ; ici, il n'y pas de déterminant article en grec, ce qui ne serait pas envisageable en français.

Le traducteur a visiblement pris plus de liberté pour traduire la seconde phrase (Figure 27). L'expression « ce mortel sur terre » n'a pas d'équivalent dans la langue source. Pour traduire βασιλέα τοῦ κόσμου, le lecteur s'attendrait à une expression du type « le souverain du monde », mais τοῦ κόσμου n'est pas traduit, etc. C'est l'*exhaustivité* de la traduction qui est ici mise à l'épreuve. Bien évidemment, il n'est pas toujours possible d'apparier un mot de la langue source à un mot de la langue cible. Dans certains cas, seules des relations de un à plusieurs ou de plusieurs à un peuvent être établies. Dans d'autres cas, des éléments présents dans une des deux langues ne sont pas présents dans l'autre. L'examen de la concordance bilingue du DSOS montre que la traduction française n'est pas satisfaisante, car elle est incomplète en de nombreux endroits et éloignée du texte source. S'il est normal que les particules δέ, καί et μέν — dont l'usage est si spécifique à la langue grecque — ne soient pas systématiquement traduites, on peut s'étonner que les mots ἀεί en 4, 32, 22, ἀκίνητος en 1, 24, 4, ἄνθρωπον en *Proe.*, 20, 11, θεασάμενος en 3, 30, 1, λευκῆς en 4, 32, 1, πόλεως en 5, 34, 7, etc. n'aient pas été pris en compte par le traducteur.

De tels alignements permettent aussi de vérifier la *cohérence* d'une traduction. L'examen parallèle du texte grec et de sa traduction française intrigue le lecteur dans quelques passages précisant les dimensions des merveilles décrites. Le Tableau 3 reprend la concordance bilingue des déterminants numériques évaluant (en stades et en coudées dans le texte grec, en pieds et en milles pour la traduction française), la hauteur et le périmètre des pyramides de Memphis, la hauteur du colosse de Rhodes, la longueur des remparts de Babylone et, enfin, leur hauteur.

GRC	FRA
La hauteur et le périmètre des pyramides de Memphis	
DSOS, 2, 26, 9 — τὸ μὲν ὕψος ἐστὶν **πήχεων τριακοσίων**, ἡ δὲ περίμετρος **σταδίων ἕξ**.	DSOS, 2, 311, 3, 2 — la hauteur est **de cinq cents pieds** et le périmètre de la base est de **trois mille six cents pieds**
Une coudée vaut 1,5 pieds. L'équivalence correcte pour la hauteur serait de 450 pieds, et non 500. Un stade vaut 600 pieds. L'équivalence du périmètre est correcte.	
La hauteur du colosse de Rhodes	
DSOS, 4, 30, 10 — ἐν ταύτῃ κολοσσὸς ἔστη **πήχεων ἑβδομήκοντα** διεσκευασμένος εἰς Ἥλιον·	DSOS, 4, 312, 1, 4 — Sur cette île se dressait un Colosse, de **cent vingt pieds** de hauteur représentant Hélios.
Une coudée vaut 1,5 pieds. L'équivalence correcte pour la hauteur serait de 105 pieds, et non 120.	
La longueur des remparts de Babylone	
DSOS, 5, 34, 6 — Βαβυλῶνα γὰρ ἐτείχισεν **τριακοσίων ἑξήκοντα σταδίων** βαλλομένη θεμελίωσιν	DSOS, 5, 314, 1, 4 — elle avait fondé, pour fortifier Babylone, des Remparts longs de **quarante et un milles**.
Un mille romain vaut 8,33 stades. L'équivalence correcte serait de 43,2 milles, et non 41.	
La hauteur des remparts de Babylone	
DSOS, 5, 34, 12 — καὶ τὸ μὲν ὕψος ἐστὶ τοῦ τείχους πλέον ἢ **πεντήκοντα πήχεων**	DSOS, 5, 314, 2, 1 — Le rempart a plus de **quatre-vingts pieds** de haut
Une coudée vaut 1,5 pieds. L'équivalence correcte serait de 75 pieds, et non 80.	

Tableau 3. Cohérence de la traduction française du DSOS : question de métrique.

Manifestement, aux yeux du traducteur, la coudée vaut une fois 1,71 pieds (120/70 pour la hauteur du colosse de Rhodes), une fois 1,67 (500/300 pour la hauteur des pyramides de Memphis), une fois 1,60 pieds (80/50 pour la hauteur des remparts de Babylone). Seul le périmètre de la base des pyramides correspond à une évaluation attendue, un stade valant 600 pieds[38]. Une « simple » traduction, sans tentative de conversion des mesures exprimées en grec, n'aurait pas nui à la compréhension du texte[39].

Les concordances bilingues sont une nouveauté du projet GRE*g*ORI. Les développements ont d'abord porté sur des alignements grec-francais (réalisés pour la mise au point des programmes informatiques utilisés à cette fin), puis sur des alignements grec-géorgien et grec-arménien. On obtient ainsi des corpus alignés, ressources très intéressantes, d'une manière générale, pour les linguistes, les traducteurs et les philologues. Aux premiers, ces outils lexicologiques fournissent d'utiles enseignements sur la langue source ou la langue cible du texte traité. Dans le cas des langues modernes, les traducteurs y recherchent, manuellement ou automatiquement, des « solutions de traduction ». Quant aux philologues, ils peuvent analyser, sur des textes anciens, quelles sont les « solutions de traductions » proposées par les traducteurs.

Une fois ces traitements effectués, les données linguistiques du texte source et du texte cible sont versées dans des bases de données communes. Les outils bilingues sont tirés de ces dernières. Les observations réalisées sur les concordances bilingues permettent de confirmer ou d'infirmer de manière tangible, sur de vastes corpus, les intuitions des philologues en matière de méthode de traduction. Les données lexicales bilingues seront

[38] Ces précisions métrologiques nous ont été transmises par le Professeur Charles Doyen (UCL) que nous remercions chaleureusement.

[39] La traduction française du DSOS pourrait avoir été produite à partir d'un autre texte source que le texte grec ; c'est une question intéressante, mais qui dépasse le cadre de cet article.

utilisées pour constituer des lexiques bilingues, outils faisant encore défaut pour l'étude des langues abordées.

Le projet GRE*g*ORI fournit également aux chercheurs différents formats d'index, monolingues ou bilingues. Les Figures 28-30 en fournissent trois exemples, respectivement un index traditionnel dit index « fin de livre », reprenant le texte du *Proemium* du DSOS, un index inverse des lemmes de la même section du DSOS et un court extrait de l'index bilingue grec-français du DSOS (dont la version intégrale est accessible sur le site du projet GRE*g*ORI)[40].

[40] Cfr note 1.

Αἴγυπτος (ἡ) { N+Top } 1
αἴγυπτον 1 *Proe.*, 20
ἀκριβής { A } 1
ἀκριβές 1 *Proe.*, 20
ἀκροατής { N+Com } 1
ἀκροατήν 1 *Proe.*, 22
Ἁλικαρνασσός { N+Top } 1
ἁλικαρνασσόν 1 *Proe.*, 20
ἀνεξάλειπτος { A } 1
ἀνεξαλείπτους 1 *Proe.*, 20
ἄνθρωπος { N+Com } 1
ἄνθρωπον 1 *Proe.*, 20
ἀνομοίως { I+Adv } 1
ἀνομοίως 1 *Proe.*, 22
ἅπαξ { I+Adv } 1
ἅπαξ 1 *Proe.*, 20
ἀποδημέω { V } 1
ἀποδημῆσαι 1 *Proe.*, 20
ἀποδημία { N+Com } 1
ἀποδημίας 1 *Proe.*, 20
ἀπολύω { V } 1
ἀπολύσασα 1 *Proe.*, 20
αὐτός { PRO+Dem } 1
αὐτός 1 *Proe.*, 22
βίος { N+Com } 1
βίος 1 *Proe.*, 20
βλέπω { V } 1
βλεπόμενα 1 *Proe.*, 22
γάρ { I+Part } 6
γάρ 6 *Proe.*, 20 (4); DSOS, *Proe.*, 22 (2)
γιγνώσκω { V } 1
γινώσκεται 1 *Proe.*, 20
δέ { I+Part } 5
δ 1 *Proe.*, 22
δέ 4 *Proe.*, 20 (4)
δείκνυμι { V } 1
δείκνυσιν 1 *Proe.*, 20
δέω (δεήσω) { V } 1
δεῖ 1 *Proe.*, 20
δή { I+Part } 1
δή 1 *Proe.*, 22
διά { I+Prep } 1
διά 1 *Proe.*, 20
διαλάμπω { V } 1
διαλάμψῃ 1 *Proe.*, 22
διαπλέω { V } 1
διαπλεῦσαι 1 *Proe.*, 20
δόξα { N+Com } 1
δόξαν 1 *Proe.*, 22
ἐάν { I+Conj } 1
ἐάν 1 *Proe.*, 20
ἐάω { V } 1
ἐᾷ 1 *Proe.*, 22
ἐγκατοπτρίζομαι { V } 1
ἐγκατοπτρισάμενος 1 *Proe.*, 20
εἴδωλον { N+Com } 1
εἰδώλων 1 *Proe.*, 20
εἰς { I+Prep } 2
εἰς 2 *Proe.*, 20 (2)
ἕκαστος { PRO+Ind } 3
ἑκάστου 1 *Proe.*, 20
ἕκαστον 2 *Proe.*, 20 (2)
ἐκλύω { V } 1
ἐκλυθέντα 1 *Proe.*, 20
Ἑλλάς { N+Top } 1
ἑλλάδος 1 *Proe.*, 20
ἐναργῶς { I+Adv } 1
ἐναργῶς 1 *Proe.*, 20
ἐνεπιδημέω { V } 1
ἐνεπιδημῆσαι 1 *Proe.*, 20
ἐνέργεια { N+Com } 1
ἐνεργείας 1 *Proe.*, 20
ἐξεργασία { N+Com } 1
ἐξεργασίας 1 *Proe.*, 20
ἔπαινος { N+Com } 1
ἐπαίνων 1 *Proe.*, 22
ἐπέρχομαι { V } 1
ἐπελθεῖν 1 *Proe.*, 20
ἐπί { I+Prep } 2
ἐπί 1 *Proe.*, 20
ἐφ΄ 1 *Proe.*, 20
ἐπιθυμία { N+Com } 1
ἐπιθυμίαν 1 *Proe.*, 20
ἐπιλανθάνομαι (–λήθω) { V } 1
ἐπιλέλησται 1 *Proe.*, 20
ἐπινεύω { V } 1
ἐπινεῦσαι 1 *Proe.*, 22
ἑπτά { NUM+Car } 2
ἑπτά 2 *Proe.*, 20 (2)
ἔργον { N+Com } 2
ἔργον 1 *Proe.*, 20
ἔργων 1 *Proe.*, 20
ἔρχομαι { V } 2
ἐλθεῖν 1 *Proe.*, 20
ἐλθών 1 *Proe.*, 20
ἔτος { N+Com } 1
ἔτεσι 1 *Proe.*, 20
Εὐφράτης { N+Top } 1
εὐφράτην 1 *Proe.*, 20
Ἔφεσος { N+Top } 1
ἔφεσον 1 *Proe.*, 20
ἐφοδεύω { V } 1
ἐφοδεύσας 1 *Proe.*, 20
ζάω { V } 1
ζῆν 1 *Proe.*, 20
Ἡλεῖος (ὁ) { N+Ant } 1
ἡλείοις 1 *Proe.*, 20
ἥλιος { N+Com } 1
ἡλίῳ 1 *Proe.*, 22
θαυμάζω { V } 2
θαυμαζόμενα 1 *Proe.*, 22
θαυμαζόμενον 1 *Proe.*, 20
θαυμαστός { A } 1
θαυμαστόν 1 *Proe.*, 20
θέαμα { N+Com } 2
θεαμάτων 2 *Proe.*, 20 (2)
θεάομαι (–άω) { V } 1
θεάσασθαι 1 *Proe.*, 20
θεωρέω { V } 1
θεωρεῖν 1 *Proe.*, 22
θεωρία { N+Com } 1
θεωρίας 1 *Proe.*, 22
ἱστορέω { V } 1
ἱστορήσας 1 *Proe.*, 20
Ἰωνία { N+Top } 1
ἰωνίας 1 *Proe.*, 20
καί { I+Part } 14
καί 14 *Proe.*, 20 (13); DSOS, *Proe.*, 22
καλέω { V } 1
καλεῖται 1 *Proe.*, 22
καλός { A } 2
καλά 1 *Proe.*, 20
καλόν 1 *Proe.*, 22
Καρία { N+Top } 1
καρίας 1 *Proe.*, 20
κατά { I+Prep } 1
κατά 1 *Proe.*, 20
κοινός { A } 1
κοινῇ 1 *Proe.*, 22
κομίζω { V } 1
κομισάμενον 1 *Proe.*, 22
κόπος { N+Com } 1
κόπῳ 1 *Proe.*, 20
κόσμος { N+Com } 1
κόσμον 1 *Proe.*, 20
λανθάνω { V } 1
λανθάνει 1 *Proe.*, 20

λέγω { V } 1
λέγω 1 *Proe.*, 20
λόγος { N+Com } 2
λόγος 1 *Proe.*, 20
λόγῳ 1 *Proe.*, 20
λοιπός { A } 1
λοιπά 1 *Proe.*, 22
μεγαλόδωρος { A } 1
μεγαλόδωρον 1 *Proe.*, 20
μέν { I+Part } 3
μέν 3 *Proe.*, 20 (2); DSOS, *Proe.*, 22
μέρος { N+Com } 1
μέρος 1 *Proe.*, 20
μνήμη { N+Com } 1
μνῆμαι 1 *Proe.*, 20
μόνος { A } 1
μόνον 1 *Proe.*, 22
ὁ { DET } 45
αἱ 1 *Proe.*, 20
ὁ 4 *Proe.*, 20 (4)
τά 4 *Proe.*, 20 (3); DSOS, *Proe.*, 22
τάς 1 *Proe.*, 20
τῇ 3 *Proe.*, 20 (2); DSOS, *Proe.*, 22
τήν 4 *Proe.*, 20 (3); DSOS, *Proe.*, 22
τῆς 8 *Proe.*, 20 (7); DSOS, *Proe.*, 22
τό 5 *Proe.*, 20 (4); DSOS, *Proe.*, 22
τοῖς 2 *Proe.*, 20 (2)
τόν 4 *Proe.*, 20 (4)
τοῦ 1 *Proe.*, 20
τούς 2 *Proe.*, 20 (2)
τῷ 1 *Proe.*, 20
τῶν 5 *Proe.*, 20 (4); DSOS, *Proe.*, 22
ὁδοιπορία { N+Com } 1
ὁδοιπορίας 1 *Proe.*, 20
οἴκοι { I+Adv } 1
οἴκοι 1 *Proe.*, 20
ὅλος { A } 1
ὅλον 1 *Proe.*, 20
ὄμμα { N+Com } 1
ὄμματα 1 *Proe.*, 20
ὁμοίως { I+Adv } 1
ὁμοίως 1 *Proe.*, 22
ὁράω { V } 3
ἑώρακεν 1 *Proe.*, 20
ἶδεν 1 *Proe.*, 20
ὁρᾶται 1 *Proe.*, 20
ὅς ἥ ὅ { PRO+Rel } 1
ὅ 1 *Proe.*, 20
ὅταν { I+Conj } 1
ὅταν 1 *Proe.*, 22
ὅτε { I+Conj } 1
ὅτε 1 *Proe.*, 20
ὅτι { I+Conj } 1
ὅτι 1 *Proe.*, 20
οὐ { I+Neg } 1
οὐκ 1 *Proe.*, 22
οὗτος { PRO+Dem } 2
ταῦτα 1 *Proe.*, 22
τοῦτο 1 Proe., 20
ὄψις (ἡ) { N+Com } 1
ὄψει 1 *Proe.*, 20
παιδεία { N+Com } 1
παιδεία 1 *Proe.*, 20
παράδοξος { A } 2
παράδοξα 1 *Proe.*, 20
παράδοξον 1 *Proe.*, 20
παραπλησίως { I+Adv } 1
παραπλησίως 1 *Proe.*, 22
παρέρχομαι { V } 1
παρελθών 1 *Proe.*, 20
παροίχομαι { V } 1
παρῳχηκεν 1 *Proe.*, 20
πᾶς { A } 1
πᾶσιν 1 *Proe.*, 20
πείθω { V } 1
πείσῃ 1 *Proe.*, 20
περί { I+Prep } 1
περί 1 *Proe.*, 20
Πέρσης (Περσίς) { A } 1
πέρσας 1 *Proe.*, 20
πιστός (πείθω) { A } 1
πιστόν 1 *Proe.*, 20
πλανάω { V } 1
πλανηθέντα 1 *Proe.*, 20
πληρόω { V } 1
πληρῶσαι 1 *Proe.*, 20
προσδίδωμι { V } 1
προσδιδοῦσα 1 *Proe.*, 20
προσηγορία { N+Com } 1
προσηγορίᾳ 1 *Proe.*, 22
προσπλέω { V } 1
προσπλεῦσαι 1 *Proe.*, 20
Ῥόδος { N+Top } 1
ῥόδῳ 1 *Proe.*, 20
σπάνιος { A } 1
σπανίοις 1 *Proe.*, 20
τέχνη { N+Com } 1
τέχνης 1 *Proe.*, 20
τόπος { N+Com } 1
τόπους 1 *Proe.*, 20
τότε { I+Adv } 1
τότε 1 *Proe.*, 20
τύπος { N+Com } 1
τύπους 1 *Proe.*, 20
φαίνω { V } 1
φανήσεται 1 *Proe.*, 20
φεύγω { V } 1
φεύγουσιν 1 *Proe.*, 20
φήμη { N+Com } 1
φήμῃ 1 *Proe.*, 20
φυλάσσω { V } 1
φυλάσσει 1 *Proe.*, 20
ψυχή { N+Com } 2
ψυχῇ 2 *Proe.*, 20 (2)

Figure 28. Index type index « fin de livre » du Proemium du DSOS

διά	1	θεωρία	1	τότε	1
ἐνέργεια	1	ἐξεργασία	1	δή	1
παιδεία	1	θέαμα	2	μνήμη	1
ἀποδημία	1	ὄμμα	1	φήμη	1
ἐπιθυμία	1	δόξα	1	τέχνη	1
Ἰωνία	1	κατά	1	ψυχή	2
Καρία	1	ἑπτά	2	καί	14
προσηγορία	1	δέ	5	θεάομαι (–άω)	1
ὁδοιπορία	1	ὅτε	1	ἐγκατοπτρίζομαι	1

ἐπιλανθάνομαι (–λήθω)	1	Ἠλεῖος (ὁ)	1	οὐ	1
παροίχομαι	1	ἥλιος	1	ἐάω	1
ἔρχομαι	2	σπάνιος	1	ζάω	1
ἐπέρχομαι	1	καλός	2	πλανάω	1
παρέρχομαι	1	ὅλος	1	ὁράω	3
δείκνυμι	1	κόσμος	1	λέγω	1
προσδίδωμι	1	ἔπαινος	1	φεύγω	1
οἴκοι	1	κοινός	1	δέω (δεήσω)	1
ἐπί	2	μόνος	1	καλέω	1
περί	1	παράδοξος	2	διαπλέω	1
ὅτι	1	λοιπός	1	προσπλέω	1
ἐάν	1	κόπος	1	ἐνεπιδημέω	1
ὅταν	1	τόπος	1	ἀποδημέω	1
μέν	3	τύπος	1	ἱστορέω	1
ἔργον	2	ἄνθρωπος	1	θεωρέω	1
εἴδωλον	1	μέρος	1	θαυμάζω	2
ἅπαξ	1	μεγαλόδωρος	1	κομίζω	1
ὁ	45	Ἔφεσος	1	πείθω	1
γάρ	6	Ἁλικαρνασσός	1	γιγνώσκω	1
Ἑλλάς	1	ἔτος	1	λανθάνω	1
πᾶς	1	ἀνεξάλειπτος	1	φαίνω	1
ἀκριβής	1	Αἴγυπτος (ἡ)	1	πληρόω	1
Πέρσης (Περσίς)	1	ἕκαστος	3	βλέπω	1
ἀκροατής	1	θαυμαστός	1	διαλάμπω	1
Εὐφράτης	1	πιστός (πείθω)	1	φυλάσσω	1
εἰς	2	αὐτός	1	ἐφοδεύω	1
ὄψις (ἡ)	1	οὗτος	2	ἐπινεύω	1
ὅς ἥ ὅ	1	ἐναργῶς	1	ἐκλύω	1
λόγος	2	ὁμοίως	1	ἀπολύω	1
Ῥόδος	1	ἀνομοίως	1		
βίος	1	παραπλησίως	1		

Figure 29. Index inverse des lemmes du Proemium du DSOS

L'index bilingue de la Figure 30 appelle quelques précisions. Pour chaque entrée, la première ligne indique le lemme grec précédé de sa fréquence dans le texte et suivi de l'indication de sa catégorie morphosyntaxique : « 1 ἀγαθός { A } ». L'utilisateur peut demander un index basé sur la succession alphabétique des lemmes grecs, comme c'est le cas ici, ou sur celle des lemmes français. La ligne suivante énumère la forme grecque dans le texte (« βελτίων »), la forme française qui lui correspond (« plus ») et enfin le lemme français accompagné de l'indication de sa catégorie morphosyntaxique (« plus { ADV } »). Les formes grecques et françaises sont également accompagnées d'une indication de leur fréquence et de leur référence dans les textes source et cible. Sous le lemme ἄδιψος, on remarque que la forme grecque ἄδιψος est traduite pas une expression française « exempte de toute soif ». Comme annoncé, l'opération d'alignement du texte sur sa traduction ne relie pas forcément un mot de la langue source à un mot de la langue cible. Les lemmes de chacune des formes françaises de l'expression sont ensuite cités (« exempt { A } de { PREP } tout { DET } soif { N } »). Plus avant dans l'index, aucune traduction n'est attribuée au lemme ἀεί, car la forme n'est pas traduite en français, comme nous l'avons déjà indiqué ci-dessus. Épinglons enfin que les deux formes attestées du lemme ἀνάβασις on été traduites différemment, l'une, ἀνάβασιν, est rendue par un infinitif (« escalader »), l'autre, ἀναβάσεως, par un nom (« ascension »).

1	**ἀγαθός { A }**							
		1	1	βελτίων DSOS, 3, 30	2	plus DSOS, 3, 312, 4	2	**plus {ADV}**
1	**ἀγέλη { N+Com }**							
		1	1	ἀγέλαις DSOS, 3, 28	1	troupeaux DSOS, 3, 312, 2	1	**troupeau {N}**
1	**ἄδιψος { A }**							
		1	1	ἄδιψος DSOS, 1, 240	1	exemptes de toute soif DSOS, 1, 310, 5	1	**exempt {A } de { PREP } tout { DET } soif { N}**
1	**ἀδύνατος { A }**							
		1	1	ἀδύνατον DSOS, 2, 24	1	impossible DSOS, 2, 311, 1	1	**impossible {A}**
1	**ἀεί { I+Adv }**							
		1	1	ἀεί DSOS, 4, 32				
1	**ἀειθαλής { A }**							
		1	1	ἀειθαλής DSOS, 1, 24	1	toujours verte DSOS, 1, 310, 4	1	**toujours {ADV } vert { A}**
1	**ἀήρ { N+Com }**							
		1	1	ἀέρι DSOS, 1, 22	1	airs DSOS, 1, 310, 1	1	**air {N}**
3	**ἀθανασία { N+Com }**							
		1	1	ἀθανασία DSOS, 3, 28	3	immortalité DSOS, 3, 312, 3	3	**immortalité {N}**
		2	2	ἀθανασίας DSOS, 6, 36	3	immortalité DSOS, 6, 314, 1	3	**immortalité {N}**
1	**ἀθάνατος { A }**							
		1	1	ἀθάνατος DSOS, 3, 28	1	immortelle DSOS, 3, 311, 1	1	**immortel {A}**
1	**Αἴγυπτος (ἡ) { N+Top }**							
		1	1	αἴγυπτον DSOS, *Proe.*, 20	1	égypte DSOS, *Intr.*, 309, 1	1	**Égypte {NPr}**
1	**Αἰθιοπικός { A }**							
		1	1	αἰθιοπική DSOS, 2, 26	2	afrique DSOS, 2, 311, 3	2	**Afrique {NPr}**
1	**αἱματίτης { A }**							
		1	1	αἱματίτης DSOS, 2, 26	1	rouge sang DSOS, 2, 311, 3	1	**rouge {A } sang { N}**
1	**αἰσχύνω { V }**							
		1	1	αἰσχύνεται DSOS, 3, 28	1	est gêné DSOS, 3, 312, 2	1	**être {V } gêner { V}**
1	**αἰτέω { V }**							
		1	1	αἰτησάμενος DSOS, 4, 30	1	demandât DSOS, 4, 312, 1	1	**demander {V}**
1	**ἀκίνητος { A }**							
		1	1	ἀκίνητος DSOS, 1, 24				
1	**ἄκμων { N+Com }**							
		1	1	ἀκμόνων DSOS, 4, 30	1	enclumes DSOS, 4, 313, 2	1	**enclume {N}**
1	**ἀκοή { N+Com }**							
		1	1	ἀκοῆς DSOS, 3, 30	1	ouï-dire DSOS, 3, 312, 4	1	**ouï-dire {N}**
1	**ἀκρεμών { N+Com }**							
		1	1	ἀκρεμόσιν DSOS, 1, 24	1	branches DSOS, 1, 310, 4	1	**branche {N}**
1	**ἀκριβής { A }**							
		1	1	ἀκριβές DSOS, *Proe.*, 20	1	détails DSOS, *Intr.*, 309, 2	1	**détail {N}**
1	**ἀκροατής { N+Com }**							
		1	1	ἀκροατήν DSOS, *Proe.*, 22	1	lecteur DSOS, *Intr.*, 309, 3	1	**lecteur {N}**
1	**Ἁλικαρνασσός { N+Top }**							
		1	1	ἁλικαρνασσόν DSOS, *Proe.*, 20	1	halicarnasse DSOS, *Intr.*, 309, 1	1	**Halicarnasse {NPr}**

2	**ἀλλά { I+Part }**							
		1	1	ἀλλ´ DSOS, 6, 36	8	mais DSOS, 6, 314, 1	8	**mais {CONJC}**
		1	1	ἀλλά DSOS, 5, 34	8	mais DSOS, 5, 314, 1	8	**mais {CONJC}**
1	**ἀλλήλων { PRO+Rec }**							
		1	1	ἀλλήλων DSOS, 1, 22	1	elles DSOS, 1, 310, 2	1	**elle {PRO}**
8	**ἄλλος { PRO+Ind }**							
		1	1	ἄλλῃ DSOS, 5, 34				
		1	2	ἄλλοις DSOS, 3, 30				
		1	2	ἄλλων DSOS, 1, 22				
		1	1	ἄλλα DSOS, 3, 28	4	autres DSOS, 3, 312, 3	4	**autre {DET}**
		1	2	ἄλλοις DSOS, 3, 28	4	autres DSOS, 3, 312, 1	4	**autre {DET}**
		2	2	ἄλλους DSOS, 4, 32 (2)	4	autres DSOS, 4, 313, 4	4	**autre {DET}**
		1	2	ἄλλων DSOS, 2, 26	2	d' autres DSOS, 2, 311, 4	2	**de {PREP } autre { DET}**
1	**Ἀλωεύς { N+Ant }**							
		1	1	ἀλωέως DSOS, 6, 36	1	aloée DSOS, 6, 314, 1	1	**Aloée {NPr}**
3	**ἀναβαίνω { V }**							
		1	1	ἀναβαίνουσι DSOS, 2, 28	1	s' élèvent DSOS, 2, 311, 5	1	**se {PRO } élever { V}**
		1	1	ἀναβάς DSOS, 4, 32	1	il atteignit DSOS, 4, 313, 6	1	**il {PRO } atteindre { V}**
		1	1	ἀναβῆναι DSOS, 4, 32	1	dut se poursuivre DSOS, 4, 313, 3	1	**devoir {V } se { PRO } poursuivre { V}**
2	**ἀνάβασις { N+Com }**							
		1	1	ἀνάβασιν DSOS, 6, 36	1	escalader DSOS, 6, 314, 1	1	**escalader {V}**
		1	1	ἀναβάσεως DSOS, 2, 26	1	ascension DSOS, 2, 311, 5	1	**ascension {N}**
1	**ἀναβιβάζω { V }**							
		1	1	ἀναβιβάζοντες DSOS, 4, 30	1	pour ériger DSOS, 4, 312, 2	1	**pour {PREP } ériger { V}**
1	**ἀνάγκη { N+Com }**							
		1	1	ἀνάγκαις DSOS, 1, 24	1	adéquats DSOS, 1, 310, 4	1	**adéquat {A}**
1	**ἀναγωγή { N+Com }**							
		1	1	ἀναγωγήν DSOS, 2, 24	1	concevoir DSOS, 2, 311, 1	1	**concevoir {V}**
1	**ἀναδείκνυμι { V }**							
		1	1	ἀνέδειξεν DSOS, 4, 30	1	fît sortir DSOS, 4, 312, 1	1	**faire {V } sortir { V}**
1	**ἀναθηλάζω { V }**							
		1	1	ἀναθηλάζει DSOS, 1, 24				

Figure 30. Index bilingue grec-français des premiers lemmes du DSOS

Tous les outils décrits ici sont générés automatiquement par des programmes spécifiques mis au point au CENTAL. Ces programmes puisent dans les bases de données les informations lexicales correspondant au type de concordance, de liste ou d'index demandés par l'utilisateur, incorporent ces données dans des formulaires qui, une fois complétés automatiquement, sont exportables dans différents formats, allant du tableur à des documents au format PDF.

8. Conclusion

Nous venons de présenter le parcours d'un texte, depuis sa numérisation jusqu'à la constitution d'outils lexicologiques bilingues.

Les (r)évolutions technologiques des dernières décennies ont profondément transformé le travail des hellénistes. Désormais, des textes et des dictionnaires sont accessibles sur Internet. Nous mentionnons ici deux réalisations, la banque de textes du Thesaurus Linguae Graecae (TLG)[41] et le lemmatiseur Eulexis[42], afin de voir si ces dernières répondent entièrement aux attentes actuelles des philologues, des linguistes ou des historiens.

L'interface du premier — qu'on ne présente plus — permet de sélectionner une œuvre sur laquelle l'utilisateur souhaite réaliser des recherches. Il est donc possible de sélectionner le texte du DSOS. L'expérience décrite ensuite est réelle. L'opérateur peut poursuivre ses investigations en demandant les formes du lemme ῥόδον/*la rose* (*via* le champ « Search for » et en activant la fonction « Lemma Search »). L'interface du TLG affiche alors les formes disponibles pour ce lemme (toujours pour le seul texte du DSOS), à savoir :

- une occurrence de ῥόδος (présentée comme forme byzantine masculine doublet du nom neutre ῥόδον) ;

- et deux occurrences de ῥόδῳ.

La fonction suivante (« Search » et « All Forms ») permet de visualiser les extraits dans lesquels s'actualisent ces trois formes. Sous l'interface du TLG, ces extraits sont identiques à ceux de la Figure 11. Il faut yajouter la citation pésentée sous la Figure 31.

[DSOS-4-00-30-7] Ῥόδος ἐστὶ πελαγία νῆσος, ἣν τὸ παλαιὸν ἐν βυθῷ κρυπτομένη

Figure 31. Occurrence de la forme Ῥόδος dans DSOS

Or, nous l'avons vu, toutes ces occurrences, y compris la dernière, désignent, dans le DSOS, le toponyme de la ville de Rhodes. L'helléniste s'attendrait à ce que les formes possibles du lemme ῥόδον/*la rose* n'apparaissent pas, puisque ce mot est absent du texte du DSOS. Les données du TLG sont donc bien lemmatisées (un dictionnaire est appliqué aux textes), mais elles ne sont pas désambiguïsées.

Le second, le lemmatiseur Eulexis, a été mis en ligne en 2014 sur site Internet du projet « Biblissima, patrimoine écrit du Moyen Âge et de la Renaissance ». Son interface présente trois champs : « Recherche de lemme », « Fléchir un lemme » et « Lemmatiser un texte grec ». Eulexis fonctionne avec des données linguistiques tirées de trois dictionnaires, le Liddell-Scott, le Pape et l'abrégé du Bailly, et les informations flexionnelles des projets Diogenes et Perseus. Le texte du DSOS a été placé dans le champ « Lemmatiser un texte grec ». L'utilisateur reçoit en réponse la liste du vocabulaire du texte. Le Tableau 4 énumère les analyses fournies par Eulexis pour les formes δεῖ et ὄψει.

[41] Cfr note 10.
[42] Cfr http ://outils.biblissima.fr/eulexis.

δεῖ	ὄψει
δέομαι lack pres ind mp 2nd sg (attic epic doric ionic) δέω1 bind pres ind mp 2nd sg (attic epic doric ionic) δέω1 bind pres imperat act 2nd sg (attic epic) δέω1 bind pres ind act 3rd sg (attic epic doric ionic) δέω1 bind imperf ind act 3rd sg (attic epic) δέω2 lack pres ind mp 2nd sg (attic epic doric ionic) δέω2 lack pres imperat act 2nd sg (attic epic) δέω2 lack pres ind act 3rd sg (attic epic doric ionic) δέω2 lack imperf ind act 3rd sg (attic epic) δεῖ there is need imperf ind act 3rd sg (attic epic ionic)	ὁράω Inscr. destombeaux des rois fut ind mid 2nd sg ὄψις aspect fem nom/voc/acc dual (attic epic) ὄψεϊ, ὄψις aspect fem dat sg (epic) ὄψις aspect fem dat sg (attic ionic) ὄψος neut nom/voc/acc dual (attic epic) ὄψεϊ, ὄψος neut dat sg (epic ionic) ὄψος neut dat sg

Tableau 4. Analyses fournies par Eulexis pour les formes δεῖ et ὄψει

Grammaticalement, toutes ces réponses sont valides. Mais rien n'indique à l'utilisateur quelle est l'analyse actualisée *in textu*.

Il y a plus interpellant. Pour la forme Ῥόδῳ, seul le lemme ῥόδον est fourni (qui ne convient pas dans le texte du DSOS), alors que le lemme toponymique, absent du Pape, apparaît bien dans le Liddell-Scott (p. 1573) et dans l'abrégé du Bailly (p. 777). La forme Ῥοδίοις est accompagnée de la mention « non-trouvé », alors que le Liddell-Scott (p. 1573) et l'abrégé du Bailly (p. 777) enregistrent bien l'un et l'autre l'entrée Ῥόδιος (que ne connaît pas le Pape). Sont également signalées comme inconnues, les formes « θεάσασθαι· » et « ⟨καὶ⟩ » (*sic*, avec le point en haut et les crochets), alors que les formes θεάσασθαι et καὶ sont, quant à elle, bien identifiées : l'examen lexical réalisé par Eulexis ne porte pas sur des formes nettoyées. Les formes « ἐπ´ » (*sic*, avec le caractère Unicode de l'apostrophe d'élision U+00B4)[43] et « α´ » (*sic*, avec le caractère Unicode de la marque numérique U+0374, pour πρῶτος/*premier*) ne sont pas non plus identifiées. Enfin, des mots plus spécifiques au vocabulaire du DSOS échappent également à l'analyse et sont eux aussi signalés comme « non trouvé » : le nom ῥαιστηροκοπίαν *s.l.* ῥαιστηροκοπία (Liddell-Scott, p. 1567), l'adjectif ἑβδομηκοντάπηχυς *s.l.* ἑβδομηκοντάπηχυς (Liddell-Scott, p. 466), l'adverbe κοχλιοειδῶς (Liddell-Scott, p. 988), etc.

Les philologues, les linguistes ou les historiens, tous les chercheurs qui ont besoin des sources, sont désormais en attente de textes entièrement analysés et désambiguïsés. Seuls des corpus parfaitement traités peuvent en effet être interrogés de manières fiables afin de servir ensuite pour d'autres travaux dans différents domaines : le lexique, la syntaxe, la création de nouvelles ressources linguistiques pour enrichir des systèmes de traitement automatique, l'extraction d'informations, etc. Grâce aux banques de textes du TLG ou du projet Perseus, les versions numériques des textes sont aujourd'hui largement accessibles aux chercheurs. Mais ces derniers sont en attente d'outils d'exploration et d'interrogation plus poussés, d'une part, mais aussi faciles à utiliser, d'autre part. À son niveau, le projet GRE*g*ORI tente d'offrir des pistes en ce domaine.

[43] Par contre, la forme « ἐπ' » (*sic*, avec l'apostrophe latine U+0027) est reconnue.

9. Bibliographie

COURTOIS, *Dictionnaires électroniques* = B. COURTOIS, *Un système de dictionnaires électroniques pour les mots simples du français*, dans *Langue Française*, 87 (1990), p. 11-22.

D.N.P. = Der Neue Pauly.

DSOS (Brodersen) = K. BRODERSEN (ed.), *Reiseführer zu den Sieben Weltwundern. Philon von Byzanz und andere antike Texte*, Francfort-sur-le-Main, Leipzig, 1992.

KINDT — PIRARD, *Couverture* = B. KINDT, M. PIRARD, *De Nazianze à Ninive. La couverture lexicale du* Dictionnaire Automatique Grec (article sous presse).

KINDT, *Principes* = B. KINDT, *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.*), dans *Byzantion*, 74 (2004), p. 213-272.

KINDT, *Traitement* = B. KINDT, *Traitement automatique de l'ambiguïté lexicale en grec ancien. Outils informatiques et ressources linguistiques*, Thèse de doctorat inédite, Louvain-la-Neuve, 2012.

LAPORTE — MONCEAUX, *Elag* = É. LAPORTE, A. MONCEAU, *Elimination of lexical ambiguities by grammars : the Elag system*, dans *Lingvisticae Investigationes*, 22 (1998-1999), p. 341-367.

LAPORTE, *Concordanciers* = É. LAPORTE, *Concordanciers et flexion automatique*, dans *Cahiers de lexicologie*, 94/1 (2009), p. 91-106.

LAPORTE, *Reduction* = É. LAPORTE, *Reduction of lexical ambiguity*, dans *Lingvisticae Investigationes*, 24/1 (2001), p. 67-103.

P.L.P. = Prosopographisches Lexikon der Palaiologenzeit, ed. E. TRAPP (et al.) (*Österreichische Akademie der Wissenschaften. Philosophisch-historische Klasse. Veröffentlichungen der Kommission für Byzantinistik*, 1, 1-12 + Addenda), Vienne, 1976-1996.

ROMER — ROMER, *Sept Merveilles* = J. ROMER, El. ROMER, *Les sept merveilles du monde*, Paris, 1996.

SILBERZTEIN, *Dictionnaires électroniques* = M. SILBERZTEIN, *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX (Informatique Linguistique)*, Paris, Milan, Barcelone, Bonn, 1993.

Thesaurus Ducae = B. KINDT, A. YANNACOPOULOU, *Thesaurus Ducae Historiae Turcobyzantinae, accedunt concordantiae Narrationis de obsidione Constantinopolitana a Ioanne Canano necnon homiliarum a Dorotheo Mitylinensi et anonymo auctore* (*Corpus Christianorum. Thesaurus Patrum Graecorum*), Turnhout, 2012.

Thesaurus Basilii = B. COULIE, B. KINDT et CETEDOC, *Thesaurus Basilii Caesariensis, Opera Omnia*. Pars I. *Introductio. Enumeratio Lemmatum et Formarum A - I*. Pars II. *Enumeratio Lemmatum et Formarum K – Ω* (*Corpus Christianorum. Thesaurus Patrum Graecorum*), Turnhout, 2002.

Thesaurus Clementis = S. DEODATI et CENTAL, *Thesaurus Clementis Alexandrini. Opera Omnia* (*Corpus Christianorum. Thesaurus Patrum Graecorum*), Turnhout, 2009.

Thesaurus Gregorii Nazianzeni = J. MOSSAY et CETEDOC, *Thesaurus Sancti Gregorii Nazianzeni*, vol. I. *Enumeratio lemmatum, Orationes, Epistulae, Testamentum* (*Corpus Christianorum. Thesaurus Patrum Graecorum*), Turnhout, 1990; J. MOSSAY, B. COULIE et CETEDOC, *Thesaurus Sancti Gregorii Nazianzeni*, vol. II. *Enumeratio lemmatum, Carmina, Christus Patiens, Vita* (*Corpus Christianorum. Thesaurus Patrum Graecorum*), Turnhout, 1991.

UNITEX 3.1beta Manuel d'Utilisation = S. PAUMIER (et al.), *UNITEX 3.1beta Manuel d'Utilisation*, Marne-la-Vallée, 2015 (état du 28 janvier 2015).

Bastien KINDT
Institut orientaliste
Place Blaise Pascal, 1
B-1348 Louvain-la-Neuve
Belgique

bastien.kindt@uclouvain.be