Processing Tools for Greek and Other Languages of the Christian Middle East

Bastien Kindt

Research Assistant – Université catholique de Louvain, Belgium

bastien.kindt@uclouvain.be

Abstract
This paper presents some *computer tools* and *linguistic resources* of the GRE*g*ORI project. These developments allow automated processing of texts written in the main languages of the Christian Middel East, such as Greek, Arabic, Syriac, Armenian and Georgian. The main goal is to provide scholars with tools (lemmatized indexes and concordances) making corpus-based linguistic information available. It focuses on the questions of text processing, lemmatization, information retrieval, and bitext alignment.

Keywords
Lemmatization; Greek; Syriac; Arabic; Armenian; Georgian; lexical tagging; POS tagging; Unitex; concordances; indexes; finite state transducers; bitext; bilingual alignment; translation memories; Unitex; mkAlign.

INTRODUCTION
The GRE*g*ORI project, carried out for many years at the Oriental Institute of the Université catholique de Louvain in Louvain-la-Neuve (Belgium) (https://uclouvain.be/fr/instituts-recherche/incal/ciol/gregori-project.html), brings together linguists, philologists and computer scientists around a common goal: developing *computer tools* and *linguistic resources* allowing automated processing of texts written in Greek [Coulie, 1996; Kindt, 2004; Kindt, 2010] and in the languages of the Christian Middle East, such as Middle Arabic [Tuerlinckx, 2004], Syriac, Armenian and Georgian [Coulie, Kindt and Pataridze, 2013]. From the point of view of computational approach all these languages still remain, especially in the case of the ancient texts, poorly resourced. These *tools* and *resources* make it possible to turn raw texts into tagged corpora enriched with lexical (lexical entry or lemma), morphosyntactical (noun, verb, adjective, etc.), and inflectional (case, gender, number, voice, tense, person, etc.) labels, defining each word-form of processed texts.

This work can lead to both monolingual and multilingual corpora. In the first case, tagged data is used in creating lemmatized concordances and indexes. It is also exploited with corpus

analysis software or data retrieval system. In the second case, the analysis is focused on translation methods. Lexical data of a *source text* is related to lexical data of a *target text* in order to highlight how lexical features from a source language (in relation with a specific cultural area) are spread and received in a target language (linked with another cultural area). In both cases, the main purpose is to provide researchers with linguistic information exclusively drawn from corpus-based evidences.

This paper focuses on the Unitex corpus processor, on the one hand, and on specific outputs such as monolingual or bilingual concordances, indexes and other lexicographical lists provided in the shape of PDF documents, on the other hand. Since two contributions included in the present issue, [Pataridze and Kindt, 2017] (*Text Alignment in Ancient Greek and Georgian: A Case-Study on the First Homily of Gregory of Nazianzus*) and [Van Elverdinghe, 2017] (*Recurrent Pattern Modelling in a Corpus of Armenian Manuscript Colophons*), are founded on the developments presented here, this paper should be regarded as a "general introduction" to these two more specific studies. Given that these contributions concern Armenian and Georgian languages, this paper deals with Greek and Syriac languages.

I. TEXT PROCESSING AND LEMMATIZING

Unitex is an open source corpus processor [Paumier, 2016 and http://unitexgramlab.org/fr]. It supports the Unicode encoding standard and is language independent. This software suite works with specific resources, such as *dictionaries* [Paumier, 2106:43-74] and *grammars* [Paumier, 2106:93-160]. *Dictionaries* provide word-forms labeled with lemmata and grammatical tags (part-of-speech tags, inflectional tags). *Grammars* are user-friendly graphical representations of lexical or syntactic patterns based on the formalism of recursive transition networks.

1.1 Unitex

The following lines illustrate how Unitex allows the user to open, to process, and to explore a Greek corpus presently composed by two hundred and forty-nine letters of Gregory of Nazianzus (329-390) published by [Gallay, 1964-1967] and [Gallay and Jourjon, 1976]. This data includes 45,971 word-forms (11,318 different forms; 4,452 lemmas). This corpus is part of a bigger corpus entitled *Corpus Epistularum Patrum Cappadocum – Corpus épistolographique des Pères cappadociens*, bringing together the letters of Basilius of Caesarea, Gregory of Nyssa, Gregory of Nazianzus, and Firmus of Caesarea, henceforth available on the website of the GRE*g*ORI project.

When user opens a corpus, the interface proposes a three-step preprocessing: 1) splitting sentences, 2) normalization of the special words or expressions, and 3) lexical lookup.

Unitex deals with the sentence as a linguistic unit. The texts are split up into sentences based on punctuation. This starting process is followed by the normalization task aimed at improving analysis of some words or expressions. In Greek, this step deals with the *crasis* or *multiword expressions* containing one lexical unit corresponding (out of context) to more than one analysis, but being non ambiguous in the given sequence.

The graphical forms of the *crasis* combine two lexical units written as a single word. The process has to provide an analysis identifying clearly the presence of these two lemmata in the single form. For instance, the forms κἀγὼ "and I" and τοὔδαφος "the ground-floor" are processed offering a lexical description including, respectively, καί "and" and ἐγὼ "I" and ὁ "the" (form τὸ) and ἔδαφος "ground-floor". After preprocessing, the word-form κἀγὼ in the sentence Πλὴν ὥσπερ ὑμεῖς κύριοι τῆς ὑμετέρας γνώμης, οὕτω **κἀγώ** τῆς ἐμῆς (Letter 153, 2)

is rewritten {@κἀγώ@,2.K} {καί,καί.I+Part} {ἐγώ,ἐγώ.PRO+Per1s} and τοὔδαφος, in the sentence κόπτω μὲν οἷον **τοὔδαφος** τοῖς ποσί (Letter 153, 2), is rewritten {@τοὔδαφος@,2.K} {τό,ὁ.DET:Nns:Ans} {ἔδαφος,ἔδαφος.N+Com}. These changes are exclusively applied to the text handled by Unitex and do not modify the original version provided by the published edition. In any case, the concordances are displaying the original Greek texts.

Among the *expressions* containing one ambiguous word without context, we can quote ὡς οἶμαι "*in my opinion*" or ὥς φησι "*as we say*". The word-form ὡς is either the conjunction ὡς "*as*" or the relatively less frequent preposition ὡς "in, by", synonymous with the more common εἰς, but exclusively used with a name of person or place such as in the sentence ἀφίκετο ὡς Περδίκκαν καὶ ἐς τὴν Χαλκιδικήν "*he arrived at Perdiccas in Chalkidiki*" [Liddell-Scott, 1958, p. 2039]. Followed by οἶμαι or φησι the word-form ὡς is unequivocally the conjunction. For these two expressions, the sentences are rewritten, after preprocessing, as κεκλήκατε ἡμᾶς ἐπὶ τὴν μητρόπολιν, **{ὡς,ὡς (ὅς).I+Conj}** {οἶμαι,οἴομαι (–ω).V}, περὶ ἐπισκόπου τι βουλευσόμενοι (Letter 43, 1) and ἀλλὰ νῦν μέν, **{ὥς,ὡς (ὅς).I+Conj}** {φησι,φημί.V} Πίνδαρος (204, 5, 147, 23). Once again, the change only concerns the text handled by Unitex.

The third step of the preprocessing involves the lexicon lookup. The GRE*g*ORI project develops its own dictionaries for each processed language. The dictionaries are Unicode text files listing the word-forms. The Greek lexicon records 441,464 word-forms connected with 67,347 lemmata and, to be used with Unitex, is converted into DELAF format [Paumier, 2016:43-46]. Each word-form is provided with a one-line lexical entry using a precise syntax: inflected form, lemma corresponding to the form, part-of-speech tag of the lemma (noun, verb, adjective, adverb, pronoun, etc.) and inflectional tags describing the inflected-form (case, gender, number, voice, tense, mood, person, etc.). For instance : τό,ὁ.DET:Nns:Ans; ἔδαφος,ἔδαφος.N+Com:Nns:Vns:Ans, etc. The DELAF formalism is also used for the other languages, such as Armenian որդիք,որդի.N+Com, Syriac ܒܪ,ܒܪܐ.NOUN or Georgian ჯაბუკი,ჯაბუკი.N+Com, these three lemmata designating "the son".

It is important to precise that the lexical lookup does not take into account the context in which the words appear. Therefore, when a word corresponds to more than one lemma, all the possibilities are memorized by the process, such as, for example, the inflected form φύσει, corresponding to both the verb φύω "engender" and the noun φύσις "nature". This is a case of lexical ambiguity (see sections 1.3 and 1.5, below).

1.2 Information retrieval

After the preprocessing, the user can search the text for words or expressions by using the "Locate Pattern" menu [Paumier, 2016:84-89]. The results are displayed in the shape of "Key Word In Context" concordances. Queries may concern inflected forms, lemmata, part-of-speech tags but also the combination of these arguments, as shown in table 1 below.

Locate Pattern	Result
εἴχομεν	The concordance of the word-from εἴχομεν
<ἔχω>	The concordance of all the inflected form of the verb ἔχω
<V>	The concordance of all the inflected form of lemmata tagged as verb
<I+Neg><ἔχω>	The concordance of all the lemmata tagged as negation followed by an inflected form of the verb ἔχω

Table 1. Samples of queries.

Fig. 1 below gives a part of the concordance corresponding to the query <I+Neg><ἔχω>. Note once again that, at this stage of the process, before lexical disambiguation, queries on both <φύω> and <φύσις> provide concordances including the word-form φύσει.

[GrNa-29-9-26-25]	εἰ δόξειε μόνος τῶν πάντων Καισάριος	**μὴ ἐσχηκέναι**	φίλους, ὁ δὴ καὶ πολλοὺς ἔχειν
[GrNa-176-1-126-18]	ἐγκαλέσομεν, οὔτε ἀργίαν, οὔτε τὸ	**μὴ ἔχειν**	ὅ τι γράψεις·{S} οὐ πρὸς σοῦ τὸ ταῦτα
[GrNa-202-14-92-11]	ἐκεῖνον τὸν ἄνωθεν ἥκοντα τὸν νοῦν	**μὴ ἔχειν,**	ἀλλὰ τὴν θεότητα τοῦ Μονογενοῦς τὴν τοῦ
[GrNa-249-19-181-14]	ὑπεροχὴ τὰ χαυνότερα ἤθη διαφυσήσῃ,	**οὐκ εἶχον**	ὅπως ἐμαυτὸν ἀτρεμεῖν νουθετήσω,
[GrNa-202-20-94-7]	τοῦ αὐτοῦ πράγματος ἀληθεῖς εἶναι φύσιν	**οὐκ ἔχει.{S}**	Πῶς οὖν ὑπέμεινέ σου ἡ μεγαλοφυὴς καὶ
[GrNa-170-1-123-5]	γραμμάτων ὑμῖν ὁμιλεῖν, ἐπειδή γε ἄλλως	**οὐκ ἔχομεν,**	οὕτως οἰκονομήσαντος {τοῦ,ὁ.DET}
[GrNa-79-7-70-13]	καὶ τὴν πνευματικὴν ἐπιστασίαν	**οὐκ ἔχοντος;{S}**	Ἐκεῖνο δέ σοι καὶ πρὸ πάντων δῆλον, ὅτι
[GrNa-96-1-79-22]	Ὑπατίῳ Πολὺν ἐζημιώμεθα χρόνον,	**οὐκ ἐχούσης**	τὸν πρῶτον ἐν ἀνδράσι τῆς πρώτης ἐν
[GrNa-82-2-72-17]	μαχομένοις αὐτοῖς ᾗ τι πλέον, τὸ ἆθλον	**οὐκ ἔχουσι**	τοῦ παλαίσματος.{S} Τῷ αὐτῷ Ἐπαινῶ ὅτι
[GrNa-31-7-28-13]	λέγω·{S} τὸ δὲ ἐζημιώθην·{S}	**οὐκ ἔχω**	οἵου σὺ χρῄζεις, τὴν Ἰλιάδα.{S} Μὴ γάρ μοι

Figure 1. Concordance of the lemmata tagged as negation and followed by an inflected form of the verb ἔχω.

1.3 Text Automaton

Unitex provides a visual representation of each sentence of the processed text, called "automaton" [Paumier, 2106:161-198], as shown on fig. 2. It begins with an arrow and ends with a square inscribed in a circle. Each word form of the sentence appears in a box. Links between boxes symbolize the continuity of the sentence. Each word-form is accompanied with the corresponding lemma and part-of-speech tag, information coming from the lexical lookup.

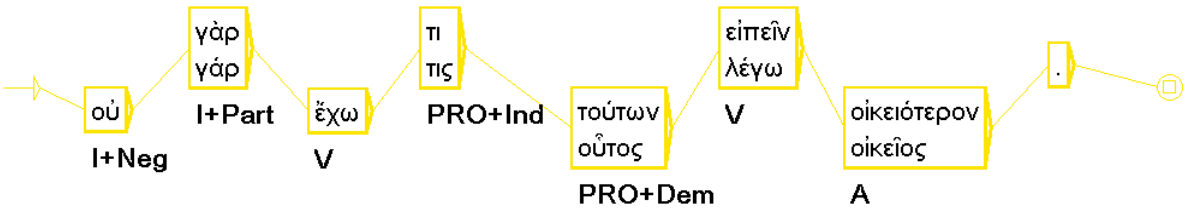

Figure 2. Automaton of the sentence *οὐ γὰρ ἔχω τι τούτων εἰπεῖν οἰκειότερον.* (Letter 1, 1).

When a word corresponds to more than one lemma, two or more boxes appear on the same horizontal axis, making different interpretations of words explicit. The automaton in fig. 3 shows the lexical ambiguity of the word-form πέτρας, inflected form of Πέτρα "Petra" (the city of Petra), or of πέτρα "stone".

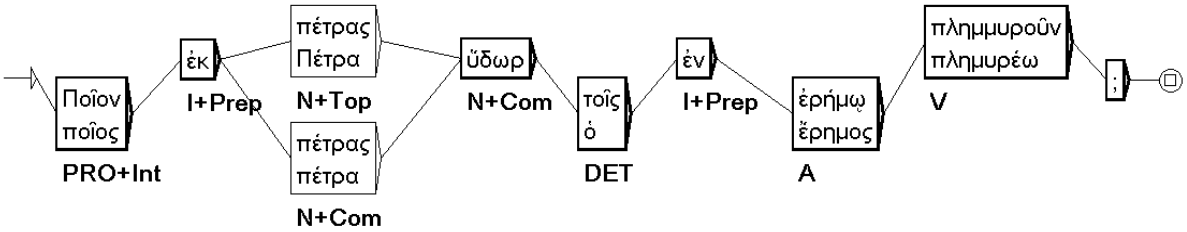

Figure 3. Automaton of the sentence *Ποῖον ἐκ πέτρας ὕδωρ τοῖς ἐν ἐρήμῳ πλημμυροῦν;* (Letter 44, 2).

The processed text is fully lemmatized, without lexical ambiguities, when the automaton shows only one box for each word of the sentences of the text, i.e. only one interpretation. At this stage, it is possible to generate valid statistics and to edit lemmatized concordances or other outputs. Lexical ambiguities are processed automatically or manually. In the second case, the user can use the lemmatization interface of Unitex, as shown in section 1.5 below.

1.4 Information retrieval by grammars

Grammars are powerful tools able to describe sequences of words. The query <I+Neg><ἔχω> used above can be graphically represented in the shape of a grammar, as shown by fig. 4. The graphical interface of Unitex allows users to draw grammars by themselves.

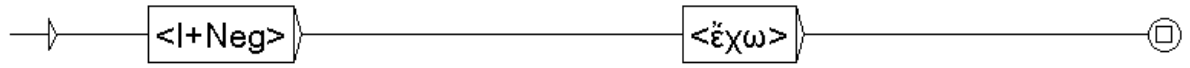

Figure 4. Grammar corresponding to the query <I+Neg><ἔχω>.

By applying this grammar to the text with Unitex, the user will be provided, once again, with the concordance of fig. 1. However, the grammars allow to search more complex patterns corresponding to a wide range of possible searches.

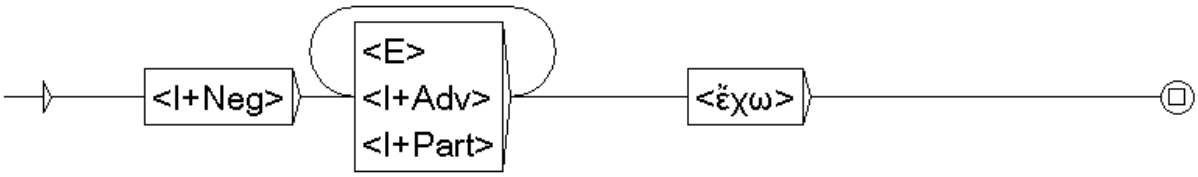

Figure 5. Grammar predicting the optional presence of an adverb or a particle between the arguments <I+Neg> and <ἔχω>.

The grammar of fig. 5 above will furnish the concordance of all the sequences constituted by a negation followed by the inflected forms of the verb ἔχω, just like the previous one, but predicts, between these two arguments, the optional presence (the sign <E> allows to recognize an empty sequence) of one or several adverbs or particles.

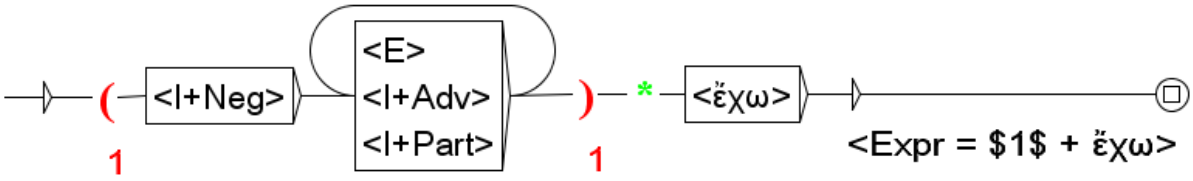

Figure 6. Grammar providing outputs and alphabetical sorting of inflected forms of the last verbal argument.

The grammars can also provide the outputs in the shape of the tags included in the results (fig. 6 above). These tags follow or surround the sequences displayed by the concordance. This grammar provides the same concordance as above, but inflected forms of ἔχω are classified according to the alphabetical order (due to the use of the asterisk before <ἔχω>). Here, the use of variables allows to repeat in the output the part of the sequence inscribed between the brackets, as shown on the concordance provided below (fig. 7).

[GrNa-249-19-181-14]	χαυνότερα ἤθη διαφυσήσῃ,	**οὐκ εἶχον <Expr=οὐκ+ἔχω>**	ὅπως ἐμαυτὸν ἀτρεμεῖν
[GrNa-29-9-26-25]	τῶν πάντων Καισάριος	**μὴ ἐσχηκέναι <Expr=μὴ+ἔχω>**	φίλους, ὁ δὴ καὶ πολλοὺς
[GrNa-81-1-72-4]	Μηδαμῶς, ὦ θαυμάσιε,	**μὴ οὕτως ἔχε <Expr=μὴ οὕτως+ἔχω>.{S}**	Τῶν μὲν γὰρ ἀκούσιος ἡ
[GrNa-202-20-94-7]	ἀληθεῖς εἶναι φύσιν	**οὐκ ἔχει <Expr=οὐκ+ἔχω>.{S}**	Πῶς οὖν ὑπέμεινέ σου ἡ
[GrNa-162-3-117-13]	Νῦν δὲ δέδοικα μὴ	**οὐχ οὕτως ἔχει <Expr=οὐχ οὕτως+ἔχω>·{S}**	μάλιστα μὲν γὰρ οὐδὲ μεμέ
[GrNa-176-1-126-18]	οὔτε ἀργίαν, οὔτε τὸ	**μὴ ἔχειν <Expr=μὴ+ἔχω>**	ὅ τι γράψεις·{S} οὐ πρὸς
[GrNa-202-14-92-11]	ἄνωθεν ἥκοντα τὸν νοῦν	**μὴ ἔχειν <Expr=μὴ+ἔχω>,**	ἀλλὰ τὴν θεότητα τοῦ
[GrNa-170-1-123-5]	ὁμιλεῖν, ἐπειδή γε ἄλλως	**οὐκ ἔχομεν <Expr=οὐκ+ἔχω>,**	οὕτως οἰκονομήσαντος
[GrNa-215-4-155-7]	ὑψηλότερον.{S} Ἡμεῖς δὲ	**οὐχ οὕτως ἔχομεν <Expr=οὐχ οὕτως+ἔχω>,**	ἀλλ᾽ ὅτι τοῦ ἑνὸς διημάρτο
[GrNa-79-7-70-13]	πνευματικὴν ἐπιστασίαν	**οὐκ ἔχοντος <Expr=οὐκ+ἔχω>;{S}**	Ἐκεῖνο δέ σοι καὶ πρὸ

Figure 7. Concordance with outputs and inflected forms of ἔχω in alphabetical order.

In fact, the preprocessing described above (section 1.1) works with this kind of grammar to analyze crasis (fig. 8 and 9) and expressions (fig. 10). In that case, the sequences are fully replaced by the content of the grammar.

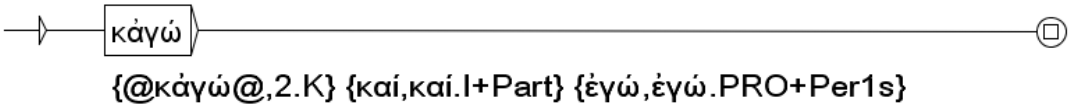

Figure 8. Grammar for the crasis κἀγώ.

Figure 9. Grammar for the crasis τοὔδαφος.

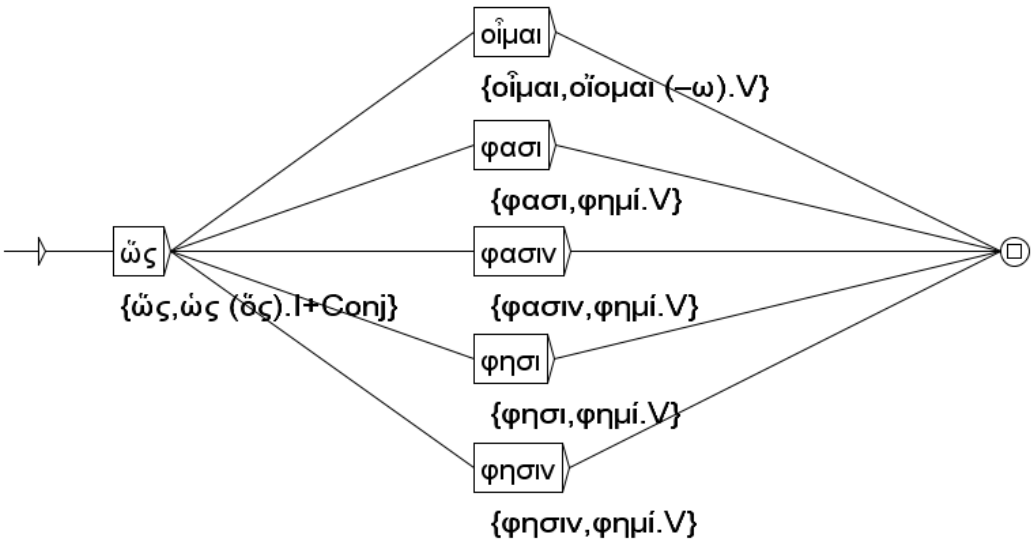

Figure 10. Grammar for the expressions ὡς οἶμαι and ὥς φησι.

Since Unitex uses the Unicode encoding and is language independent, as written above, this kind of grammars can be adapted to analyse similar linguistic facts in Armenian or Georgian. As regards Armenian, a case study is provided by [Van Elverdinghe, 2017], in this issue.

1.5 Lemmatization

Unitex provides a lemmatization interface (fig. 11) allowing to resolve lexical ambiguities manually. A special window displays the concordance and the text automaton of sentences containing lexical ambiguities.

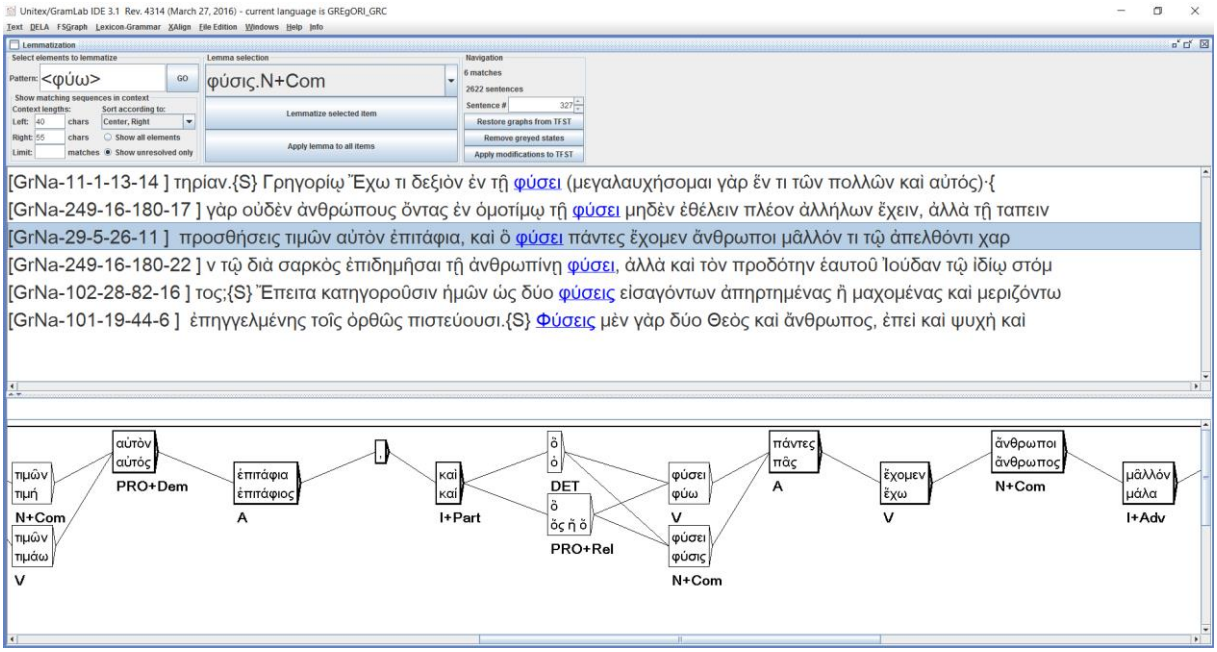

Figure 11. Lemmatization interface of Unitex.

By clicking on superfluous boxes – for instance, boxes with lemma "ὁ.DET" and lemma "φύω.V", the user can remove inadequate interpretations and solve ambiguities in order to obtain a fully lemmatized and disambiguated text, as shown in fig. 12.

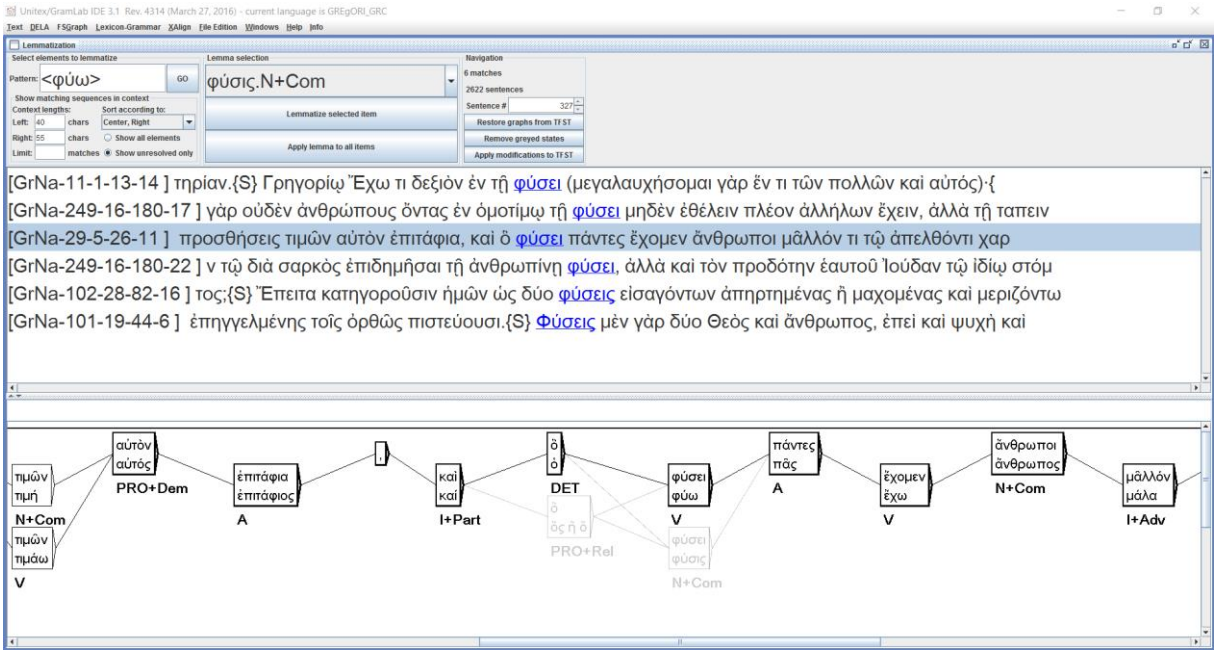

Figure 12. Lemmatization interface with elimination of useless boxes.

All linguistic resources, dictionaries and grammars, can be created, updated and corrected by the users, without any special knowledge of computer programming. The linguistic data is not encapsulated in a "black box" and remains accessible and readable.

At the end of the process, all lemmatized and disambiguated data are exported and gathered in a relational database system. Then, the publishing tools (see section III) allow to edit the special outputs in PDF format, such as concordances, indexes and other lexicographical lists.

II. TEXT ALIGMENT

The GRE*g*ORI project processes texts in Greek, Arabic, Syriac, Armenian and Georgian using appropriate resources for each language. In this context, we are interested in comparing the lexical data of a *source text* with its ancient or modern translations, the *target text*, in order to provide scholars with corpus-based information about translation methods. This task requires alignment software such as mkAlign [Fleury, 2012 and http://www.tal.univ-paris3.fr/mkAlign/]. The mkAlign software allows users to pair words or expressions of a source text with words or expressions of a target text.

The mkAlign interface is divided into two columns. The user loads the source text in the left column and the target text in the right one. The first alignment is automatically provided on the basis of the punctuation or textual references of the edition (chapters, paragraphs, etc.). The sample provided here concerns the Greek text of the Homily 13 by Gregory of Nazianzus, aligned with its Syriac (S2) version [Schmidt, 2002]. Then, the user can move the words or expressions to align the relevant translation units. Word by word alignment is not always possible, as shown on fig. 13, below, where the Greek form αἱρόμεναι is aligned with three Syriac words. Aligned data are afterwards exported as bitext files in the Translation Memory eXchange format (.tmx) and gathered in the relational database system. The bitext alignment is currently done manually. In the near future, the bilingual data will be used as translation memories in order to automatize the alignment for recurring translation units.

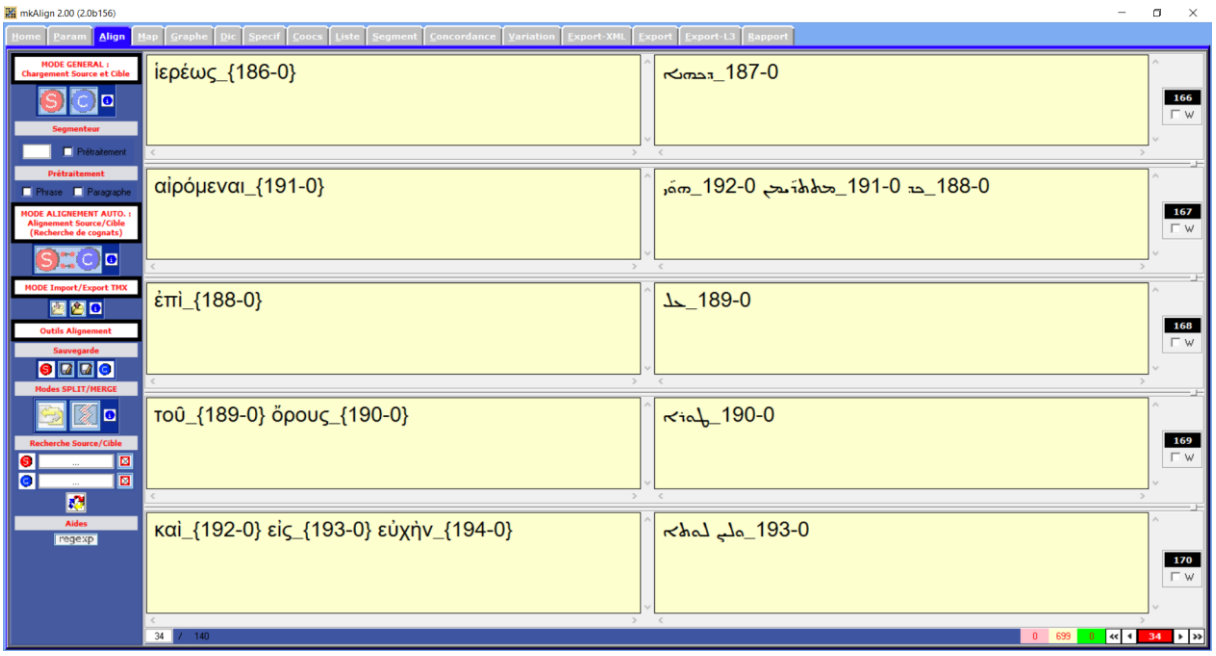

Figure 13. The mkAlign interface.

III. CONCORDANCES, INDEXES AND LEXICOGRAPHICAL LISTS

Specific tools are developed by the Centre de traitement automatique du langage (Cental) at the UCL (https://uclouvain.be/fr/instituts-recherche/ilc/cental) in order to edit concordances, indexes and other types of lexicographical lists. The purpose of these developments is to extract

the lexical data out of the databases and inject them in the frameworks corresponding to the output required by the user.

Fig. 14 shows the lemmatized concordance of the verbs included in the letters of Gregory of Nazianzus. In this PDF document, lemmata are displayed in green (accompanied with a frequency rate), POS tags in red and word-forms in blue, with downstream and upstream contexts. The references to the edition appear on the left of the downstream context. The bookmarks list lemmata and allow browsing the document.

Fig. 15 gives a sample of the alphabetical list of these verbs and Fig. 16 lists these verbs according to their frequency in the corpus.

Fig. 17 offers a sample of a bilingual Greek-Syriac concordance of Homily 13 by Gregory of Nazianzus. This document displays all the lexical data both for the source text and for the target text, including roots (in red) for Syriac. Such a bilingual concordance is the so called “scholar concordance”. Fig. 18 provides the lexical data (lemmata and POS tags) only for the source text. Such a bilingual concordance is a so-called “light concordance”, being more easily readable for the user.

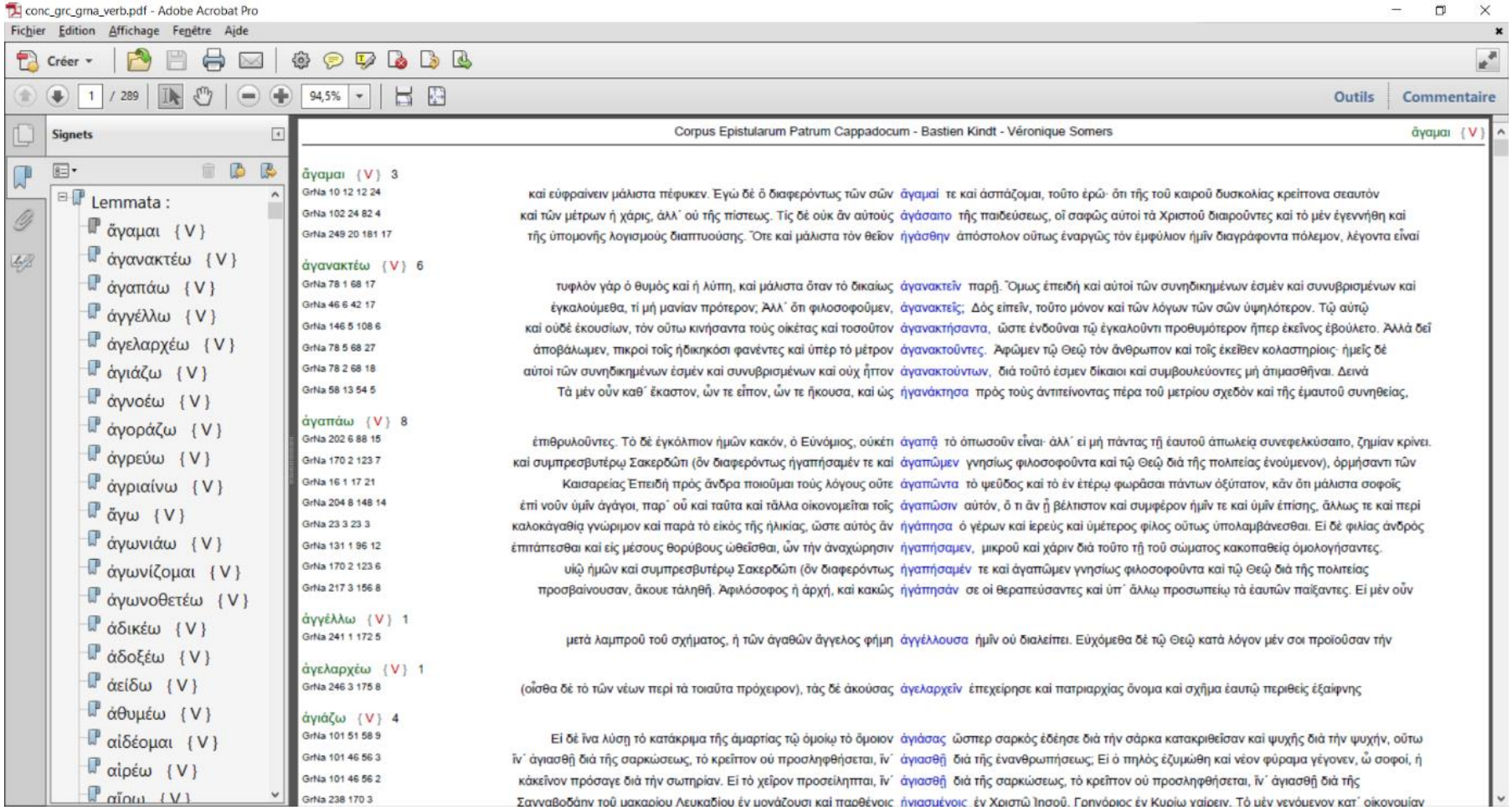

Fig. 14. Lemmatized concordance of the verbs in the corpus of the letters by Gregory of Nazianzus

Corpus Epistularum Patrum Cappadocum - Bastien Kindt - Véronique Somers

ἄγαμαι {V}	3	ἀληθεύω {V}	1	ἀνακεράννυμι {V}	1	ἀνεγείρω {V}	1
ἀγανακτέω {V}	6	ἁλιεύω {V}	1	ἀνακομίζω {V}	1	ἀνέλκω {V}	2
ἀγαπάω {V}	8	ἁλίσκομαι {V}	2	ἀνακόπτω {V}	3	ἀνερωτάω {V}	1
ἀγγέλλω {V}	1	ἀλληγορέω {V}	1	ἀνακύπτω {V}	1	ἀνευρίσκω {V}	3
ἀγελαρχέω {V}	1	ἀλλοτριόω {V}	1	ἀναλαμβάνω {V}	5	ἀνέχω {V}	11
ἀγιάζω {V}	4	ἁμαρτάνω {V}	7	ἀναλίσκω {V}	1	ἀνηβάσκω {V}	1
ἀγνοέω {V}	19	ἁμάω (θερίζω) {V}	1	ἀναλύω {V}	3	ἀνήκω {V}	1
ἀγοράζω {V}	1	ἀμείβω {V}	5	ἀναμένω {V}	6	ἀνθέω {V}	1
ἀγρεύω {V}	2	ἀμελέω {V}	7	ἀναμίγνυμι {V}	1	ἀνιάω {V}	7
ἀγριαίνω {V}	1	ἀμνημονέω {V}	1	ἀναμιμνήσκω {V}	3	ἀνίημι {V}	5
ἄγω {V}	29	ἀμύνω {V}	2	ἀνανεόομαι (-όω) {V}	1	ἀνίστημι {V}	3
ἀγωνιάω {V}	2	ἀμφιέννυμι {V}	1	ἀνανεύω {V}	1	ἀνίσχω {V}	1
ἀγωνίζομαι {V}	14	ἀμφισβητέω {V}	1	ἀναπαύω {V}	7	ἀνοηταίνω {V}	1
ἀγωνοθετέω {V}	1	ἀναβαίνω {V}	3	ἀναπείθω {V}	1	ἀνοίγω {V}	1
ἀδικέω {V}	34	ἀναβάλλω {V}	1	ἀναπλάσσω {V}	1	ἀνομμώζω {V}	1
ἀδοξέω {V}	1	ἀναγιγνώσκω {V}	4	ἀναπληρόω {V}	5	ἀντεγείρω {V}	1
ἀείδω {V}	11	ἀναγκάζω {V}	2	ἀναπνέω {V}	4	ἀντεισέρχομαι {V}	1
ἀθυμέω {V}	1	ἀναγράφω {V}	1	ἀνάπτω {V}	4	ἀντεισφέρω {V}	1
αἰδέομαι {V}	27	ἀνάγω {V}	2	ἀνασείω {V}	1	ἀντεξάγω {V}	1
αἱρέω {V}	5	ἀναδείκνυμι {V}	1	ἀναστατόω {V}	1	ἀντεπιδείκνυμι {V}	1
αἴρω {V}	5	ἀναδέχομαι {V}	1	ἀνατείνω {V}	1	ἀντεπιτίθημι {V}	1
αἰσθάνομαι {V}	9	ἀναδιδάσκω {V}	2	ἀνατίθημι {V}	2	ἀντερῶ {V}	1
αἰσχύνω {V}	22	ἀναδύομαι {V}	3	ἀνατλῆναι {V}	1	ἀντέχω {V}	3
αἰτέω {V}	25	ἀναζητέω {V}	1	ἀνατρέπω {V}	1	ἀντιβαίνω {V}	1
αἰτιάομαι {V}	9	ἀναθάλπω {V}	1	ἀνατρέφω {V}	1	ἀντιβλέπω {V}	1
ἀκηδιάω {V}	1	ἀναίνομαι {V}	1	ἀναφέρω {V}	2	ἀντιδίδωμι {V}	8
ἀκολουθέω {V}	3	ἀναιρέω {V}	2	ἀναφύω {V}	1	ἀντιλαμβάνω {V}	2
ἀκούω {V}	32	ἀνακαθαίρω {V}	1	ἀναχέω {V}	1	ἀντιλέγω {V}	6
ἀκροάομαι {V}	1	ἀνακαινίζω {V}	2	ἀναχωρέω {V}	3	ἀντιξύω {V}	1
ἀλγέω {V}	15	ἀνακαλέω {V}	3	ἀναψύχω {V}	2	ἀντισόω {V}	1
ἀλείφω {V}	2	ἀνακαλύπτω {V}	1	ἀνδρίζω {V}	3	ἀντισταθμέω {V}	1

21-11-17 Liste 'Lemmes - Analyses' Page 1 sur 13

Fig. 15. Alphabetial index of the verbs in the corpus of the letters by Gregory of Nazianzus

Corpus Epistularum Patrum Cappadocum - Bastien Kindt - Véronique Somers

εἰμί {V}	462	εὑρίσκω {V}	35	γνωρίζω {V}	22	φθέγγομαι {V}	16
ἔχω {V}	286	ἀδικέω {V}	34	μιμνήσκω {V}	22	φρονέω {V}	16
λέγω {V}	219	πράσσω {V}	34	παρίστημι {V}	22	ἀλγέω {V}	15
γίγνομαι {V}	180	ἐπιστέλλω {V}	33	χρή {V}	22	ἀπολαύω {V}	15
ποιέω {V}	176	ἀκούω {V}	32	θεραπεύω {V}	21	εὐφραίνω {V}	15
οἶδα {V}	129	κρίνω {V}	32	ὑπολαμβάνω {V}	21	μέλλω {V}	15
δέω (δεήσω) {V}	128	θαυμάζω {V}	31	κινέω {V}	20	ὁμολογέω {V}	15
γράφω {V}	104	λυπέω {V}	31	κρατέω {V}	20	σιωπάω {V}	15
δίδωμι {V}	85	ἀξιόω {V}	30	ποθέω {V}	20	φεύγω {V}	15
ὁράω {V}	78	ἄγω {V}	29	φυλάσσω {V}	20	ἀγωνίζομαι {V}	14
πάσχω {V}	73	ἄρχω {V}	29	ἀγνοέω {V}	19	ἐάω {V}	14
βούλομαι {V}	68	χράομαι (-άω) {V}	29	ἀπαιτέω {V}	19	κελεύω {V}	14
πάρειμι (εἰμί) {V}	67	ἥκω {V}	28	ἔρχομαι {V}	19	μένω {V}	14
πείθω {V}	65	παρακαλέω {V}	28	ζάω {V}	19	παρέχω {V}	14
δοκέω {V}	62	πιστεύω {V}	28	κάμνω {V}	19	σῴζω {V}	14
φημί {V}	60	αἰδέομαι {V}	27	λύω {V}	19	ἀπατάω {V}	13
τιμάω {V}	55	ἐθέλω {V}	27	εὔχομαι {V}	18	ἀφίημι {V}	13
τυγχάνω {V}	55	ἐρῶ {V}	27	ζημιόω {V}	18	διαλέγω {V}	13
φέρω {V}	52	διδάσκω {V}	26	μανθάνω {V}	18	καταγιγνώσκω {V}	13
δέχομαι {V}	51	αἰτέω {V}	25	βοηθέω {V}	17	καταλείπω {V}	13
ἐπαινέω {V}	50	ἐγκαλέω {V}	25	δηλόω {V}	17	κατέχω {V}	13
δύναμαι {V}	47	σπουδάζω {V}	25	ἐλπίζω {V}	17	μέλω {V}	13
νομίζω {V}	47	χαίρω {V}	25	ἡσσάομαι (-άω) {V}	17	μετέχω {V}	13
χαρίζομαι (-ω) {V}	47	βουλεύω {V}	24	ἵστημι {V}	17	μιμέομαι {V}	13
γιγνώσκω {V}	46	ἔοικα {V}	24	καταλύω {V}	17	ὀνομάζω {V}	13
φαίνω {V}	46	προσάγω {V}	24	καταξιόω {V}	17	ὁρμάω {V}	13
φιλοσοφέω {V}	46	καλέω {V}	23	τίθημι {V}	17	πληρόω {V}	13
λαμβάνω {V}	44	νικάω {V}	23	φροντίζω {V}	17	προσαγορεύω {V}	13
θαρσέω {V}	43	συγχωρέω {V}	23	παίζω {V}	16	προστρέχω {V}	13
οἴομαι (-ω) {V}	42	αἰσχύνω {V}	22	πρεσβεύω {V}	16	συγγιγνώσκω {V}	13
προστίθημι {V}	40	βλέπω {V}	22	πυνθάνομαι {V}	16	συνάγω {V}	13

21-11-17 Liste 'Lemmes - Analyses' Page 1 sur 13

Fig. 16. Frequency index of the verbs in the corpus of the letters by Gregory of Nazianzus

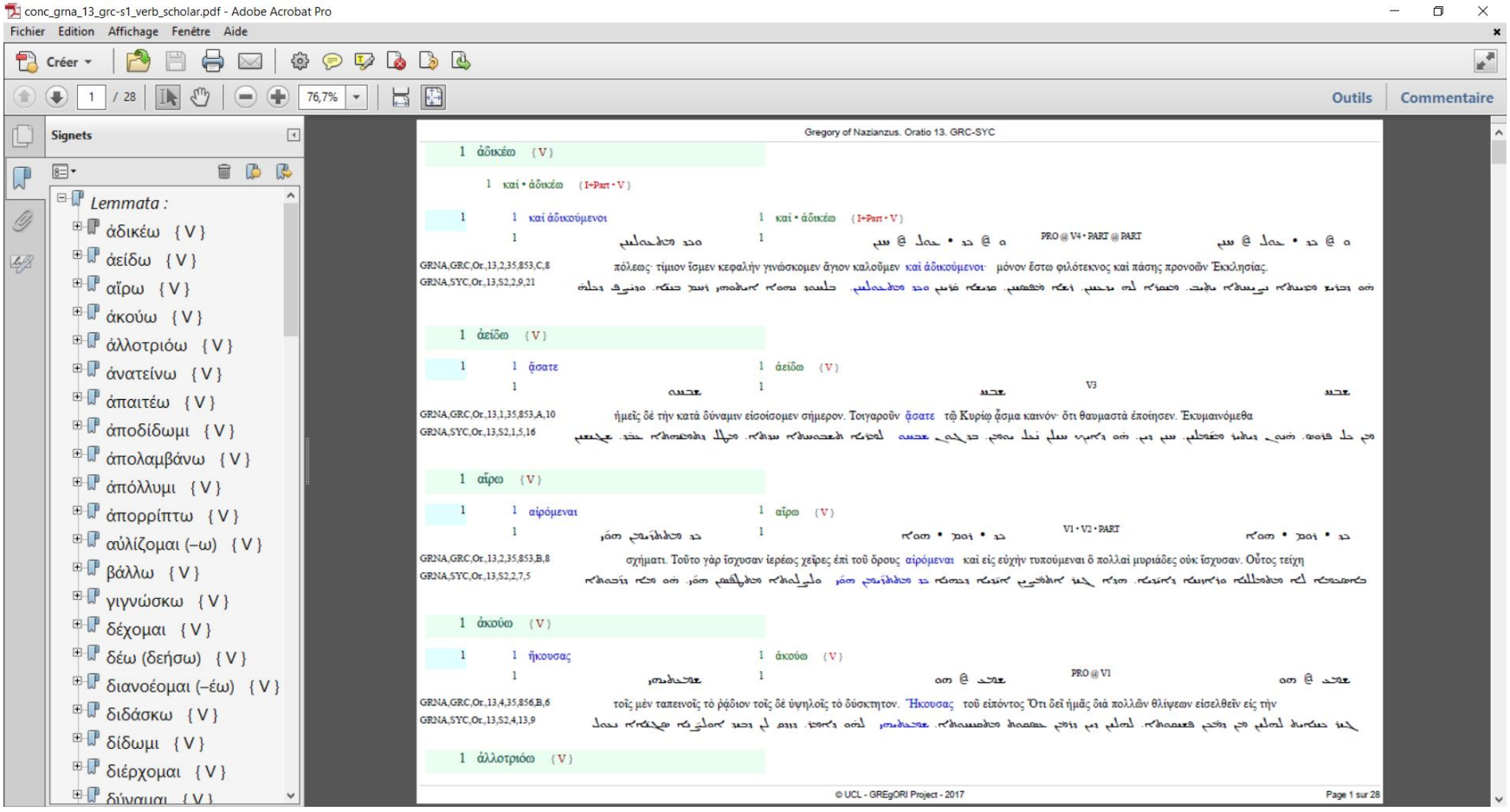

Fig. 17. Bilingual Greek-Syriac (S2) concordance of the 13th homily by Gregory of Nazianzus (scholar concordance)

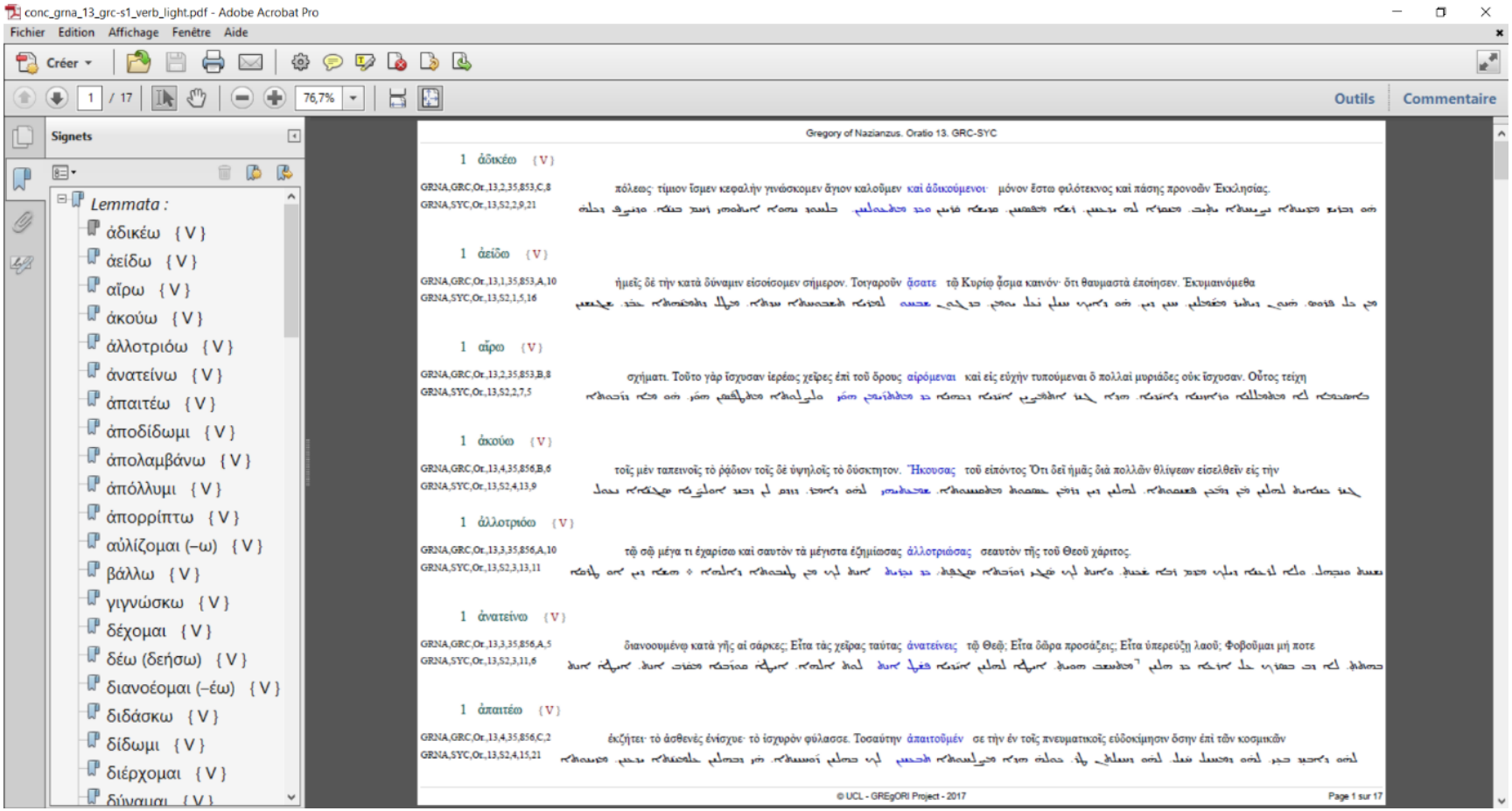

Fig. 18. Bilingual Greek-Syriac (S2) concordance of the 13th homily by Gregory of Nazianzus (light concordance)

CONCLUSION

For many years, all these technical developments and linguistic data are used to provide scholars with lemmatized concordances. These concordances were formerly published by Brepols Publishers on microfiches in the *Thesaurus Patrum Graecorum* series (see [Coulie, 1996; Kindt, 2010]. This use of microfiches is however rightly characterized as a "doubtful aspect" since [Trapp, 2008:97]. Now concordances and others lexical outputs are directly available on the website of the GRE*g*ORI project.

Some collaborators made use of these computerized resources to build lemmatized indexes (*index nominum*, *index verborum*, etc.) included in their editions of ancient texts (see for instance [Schmidt, 2001], [Auwers, 2005], [Sanpeur, 2007], [Amato, 2009], [Auwers, 2011] and more recently [Calzolari, 2017]). The GRE*g*ORI project also takes part in international research programs, such as the *Syriac Galen Palimpsest* project led by the University of Manchester or the study undertaken at the University of Lausanne and devoted to the comparative analysis of the *Iliad* and its Byzantine translation contained in the manuscript Genavensis 44.

To date, a web-based interface allowing to consult lemmatized concordances on personal computers via Internet, is being prepared in collaboration with the Cental. A beta-version of this interface is used by Peeters Publishers (Leuven) in order to pave the path for a forthcoming digital version of the well-known *Corpus Scriptorum Christianorum Orientalium*, the series dealing with some of the most important ancient sources written in languages of the Christian Middle East.

References

Amato E. Severus Sophista Alexandrinus Progymnasmata quae exstant omnia […] accedunt Callinici Petraei et Adriani Tyrii sophistarum testimonia et fragmenta necnon incerti auctoris ethopoeia nondum vulgata. Bibliotheca Scriptorum Graecorum Romanorum Teubneriana, Walter de Gruyter (Berlin, New York), 2009.

Auwers J.-M. Concordance du Siracide (Grec II et Sacra Parallela), avec la collaboration d'Églantine Proksch-Strajtmann. Cahiers de la Revue Biblique, 58. Gabalda (Paris), 2005.

Auwers J.-M. Procopii Gazaei Epitome in Canticum Canticorum, cum praefatione a Jean-Marie Auwers et Marie-Gabrielle Guérard curata. Corpus Christianorum. Series Graeca, 67. Brepols Publishers (Turnhout), 2011.

Calzolari, V. Apocrypha Armeniaca. Acta Pauli et Theclae, Prodigia Theclae, Martyrium Pauli. Corpus Christianorum. Series Apocryphorum, 20. Brepols Publishers (Turnhout), 2017.

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

Coulie B., Kindt B. and Pataridze T. Lemmatisation automatique des sources en géorgien ancien. *Le Muséon*. 2013;126:161-201.

Fleury S. mkAlign, Manuel d'utilisation, Université Sorbonne Nouvelle Paris 3, juin 2012. [get PDF]

Gallay P. and Jourjon M. Lettres théologiques. Sources Chrétiennes, 280. Les Éditions du Cerf (Paris),1976.

Gallay P. Saint Grégoire de Nazianze, Lettres. Collection des Universités de France. Les Belles Lettres (Paris), 1964-1967.

Kindt B. Des concordances et du *Thesaurus Patrum Graecorum* : la ligne éditoriale du Projet de recherche en lexicologie grecque (1990-2010). Schmidt A. *Studia Nazianzenica* II. Corpus Christianorum. Series Graeca. 73. Corpus Nazianzenum, 24. Brepols Publishers (Turnhout), 2010:43-120.

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

Liddell H.G., Scott R. and Jones H.S. A Greek-English Lexicon, 9th ed, Oxford University Press. Clarendon Press (Oxford), 1940, repr. 1977 with Barber E. Greek-English Lexicon. A Supplement, Oxford University Press. Clarendon Press (Oxford), 1968; see also Glare G. W. Greek-English Lexicon. Revised Supplement, Oxford University Press. Clarendon Press (Oxford), 1996.

Pataridze T. and Kindt B. Text Alignment in Ancient Greek and Georgian: A Case-Study on the First Homily of Gregory of Nazianzus. *Journal of Data Mining and Digital Humanities*. 2017;Special Issue.

Paumier S. Unitex 3.1 User Manual, Université Paris-Est Marne-la-Vallée, March 27, 2016. [get PDF]

Sanspeur Cl. Sancti Gregorii Nazianzeni Opera. Versio armeniaca, IV. Oratio VI. Corpus Christianorum. Series Graeca, 61. Corpus Nazianzenum, 21. Brepols Publishers (Turnhout), 2007.

Schmidt A.B. Sancti Gregorii Nazianzeni Opera. Versio Syriaca. II. Orationes XIII, XLI. Corpus Christianorum. Series Graeca, 47 ; Corpus Naizanzenum, 15. Brepols Publishers (Turnhout, Leuven), 2002.
Schmidt Th. Basilii Minimi in Gregorii Nazianzeni Orationem xxxviii Commentarii. Corpus Christianorum. Series Graeca, 46. Corpus Nazianzenum. 13. Brepols Publishers (Turnhout, Leuven), 2001.
Trapp E. Lexicography and electronic textual resources. Jeffreys E., Haldon J.F., Cormack R. (eds) The Oxford Handbook of Byzantine Studies. Oxford University Press (Oxford), 2008:95-100.
Tuerlinckx L. La lemmatisation de l'arabe non classique. Dister A., Fairon C. and Purnelle G. (eds) *Le poids des mots. 7es Journées internationales d'Analyse statistique des Données Textuelles, 10-12 mars 2004, Louvain-la-Neuve*. Vol 2. Cahiers du Cental, II. Presses Universitaires de Louvain (Louvain-la-Neuve), 2004:1069-1078. [get PDF]
Van Elverdinghe E. Recurrent Pattern Modelling in a Corpus of Armenian Manuscript Colophons. *Journal of Data Mining and Digital Humanities*. 2017;Special Issue.