A Tool for Multi-Word Expression Extraction in Modern Greek Using Syntactic Parsing

Athina Michou
University of Geneva
Geneva, Switzerland
Athina.Michou@unige.ch

Violeta Seretan
University of Geneva
Geneva, Switzerland
Violeta.Seretan@unige.ch

Abstract

This paper presents a tool for extracting multi-word expressions from corpora in Modern Greek, which is used together with a parallel concordancer to augment the lexicon of a rule-based machine-translation system. The tool is part of a larger extraction system that relies, in turn, on a multilingual parser developed over the past decade in our laboratory. The paper reviews the various NLP modules and resources which enable the retrieval of Greek multi-word expressions and their translations: the Greek parser, its lexical database, the extraction and concordancing system.

1 Introduction

In today's multilingual society, there is a pressing need for building translation resources, such as large-coverage multilingual lexicons, translation systems or translation aid tools, especially due to the increasing interest in computer-assisted translation.

This paper presents a tool intended to assist translators/lexicographers dealing with Greek[1] as a source or a target language. The tool deals specifically with multi-lexeme lexical items, also called *multi-word expressions* (henceforth MWEs). Its main functionalities are: 1) the robust parsing of Greek text corpora and the syntax-based detection of word combinations that are likely to constitute MWEs, and 2) concordance and alignment functions supporting the manual creation of monolingual and bilingual MWE lexicons.

The tool relies on a symbolic parsing technology, and is part of FipsCo, a larger extraction system (Seretan, 2008) which has previously been used to build MWE resources for other languages, including English, French, Spanish, and Italian. Its extension to Greek will ultimately enable the inclusion of this language in the list of languages supported by an in-house translation system.

The paper is structured as follows. Section 2 introduces the Greek parser and its lexical database. Section 3 provides a description of Greek MWEs, including a syntactic classification for these. Section 4 presents the extraction tool, and Section 5 concludes the paper.

2 The Greek parser

The Greek parser is part of Fips, a multilingual symbolic parser that deals, among other languages, with English, French, Spanish, Italian, and German (Wehrli, 2007). The Greek version, FipsGreek (Michou, 2007), has recently reached an acceptable level of lexical and grammatical coverage.

Fips relies on generative grammar concepts, and is basically made up of a generic parsing module which can be refined in order to suit the specific needs of a particular language. Currently, there are approximately 60 grammar rules defined for Greek, allowing for the complete parse of about 50% of the sentences in a corpus like Europarl (Koehn, 2005), which contains proceedings of the European Parliament. For the remaining sentences, partial analyses are instead proposed for the chunks identified.

One of the key components of the parser is its (manually-built) lexicon. It contains detailed morphosyntactic and semantic information, namely, selectional properties, subcategorization information, and syntactico-semantic features that are likely to influence the syntactic analysis.

The Greek monolingual lexicon presently contains about 110000 words corresponding to 16000

[1] For the sake of simplicity, we will henceforth use the term Greek to refer to Modern Greek.

lexemes,[2] and a limited number of MWEs (about 500). The bilingual lexicon used by our translation system contains slightly more than 8000 Greek-French/French-Greek equivalents.

3 MWEs in Modern Greek

Greek is a language which exhibits a high MWE productivity, with new compound words being created especially in the science and technology domains. Sometimes, existing words are transformed in order to denote new concepts; also, numerous neologisms are created or borrowed from other languages.

A frequent type of multi-word constructions in Greek are special noun phrases, called *lexical phrases* (Anastasiadi-Symeonidi, 1986) or *loose multi-word compounds* (Ralli, 2005):

- Adjective+Noun: ανοιχτή θάλασσα (anichti thalassa, 'open sea'), παιδική χαρά (pediki chara, 'kindergarten');
- Noun+Noun$_{GEN}$: ζώνη ασφαλείας (zoni asfalias, 'safety belt'), φόρος εισοδήματος (foros isodimatos, 'income tax');
- Noun+Noun$_{NOM}$ (head-complement relation): παιδί-θαύμα (pedi-thavma, 'child prodigy'), συζήτηση-μαραθώνιος (syzitisi-marathonios, 'marathon talks') ;
- Noun$_{NOM}$+Noun$_{NOM}$ (coordination relation): καναπές-κρεβάτι (kanapes-krevati, 'sofa bed'), γιατρός-νοσοκόμος (yiatros-nosokomos, 'doctor-nurse').

A large body of Greek MWEs constitute *collocations* (typical word associations whose meaning is easy to decode, but whose component items are difficult to predict), such as καταρρίπτω ένα ρεκόρ (kataripto ena rekor, 'to break a record'), in which the verbal collocate καταρρίπτω ('shake down') is unpredictable. Collocations may occur in a wide range of syntactic types. Some of the configurations taken into account in our work are:

- Noun(Subject)+Verb: η συζήτηση λήγει (i sizitisi liyi, 'discussion ends');
- Adjective+Noun: θανατική ποινή (thanatiki pini, 'death penalty');
- Verb+Noun(Object): διατρέχω κίνδυνο (diatrecho kindino, 'run a risk');
- Verb+Preposition+Noun(Argument): καταδικάζω σε θάνατο (katadikazo se thanato, 'to sentence to death');
- Verb+Preposition: προσανατολίζομαι προς (prosanatolizome pros, 'to orient to');
- Noun+Preposition+Noun: προτροπή για ανάπτυξη (protropi yia anaptiksi, 'incitement to development');
- Preposition+Noun: υπό συζήτηση (ipo sizitisi, 'under discussion');
- Verb+Adverb: χειροκροτώ θερμά (xirokroto therma, 'applause warmly');
- Adverb+Adjective: γενετικά τροποποιημένος (yenetika tropopiimenos, 'genetically modified');
- Adjective+Preposition: εξαρτημένος από (eksartimenos apo, 'dependent on').

In addition, Greek MWEs cover other types of constructions, such as:

- one-word compounds: ερυθρόδερμος (erithrodermos, 'red skin'), λυκόσκυλο (likoskylo, 'wolfhound');
- adverbial phrases: εκ των προτέρων (ek ton proteron, 'a priori, in principle');
- idiomatic expressions (whose meaning is difficult to decode): γίνομαι χαλί να με πατήσεις (yinome xali na me patisis, literally, *become a carpet to walk all over*; 'be ready to satisfy any wish').

4 The MWE Extraction Tool

MWEs constitute a high proportion of the lexicon of a language, and are crucial for many NLP tasks (Sag et al., 2002). This section introduces the tool we developed for augmenting the coverage of our monolingual/bilingual MWE lexicons.

4.1 Extraction

As we already mentioned, the Greek MWE extractor is part of FipsCo, a larger extraction system based on a symbolic parsing technology (Seretan, 2008) which we previously applied on text corpora in other languages. The recent development of the Greek parser enabled us to extend it and apply it to Greek.

[2] Most of the inflected forms were automatically obtained through morphological generation; that is, the base word was combined with the appropriate suffixes, according to a given inflectional paradigm. A number of 25 inflection classes have been defined for Greek nouns, 11 for verbs, and 10 for adjectives.

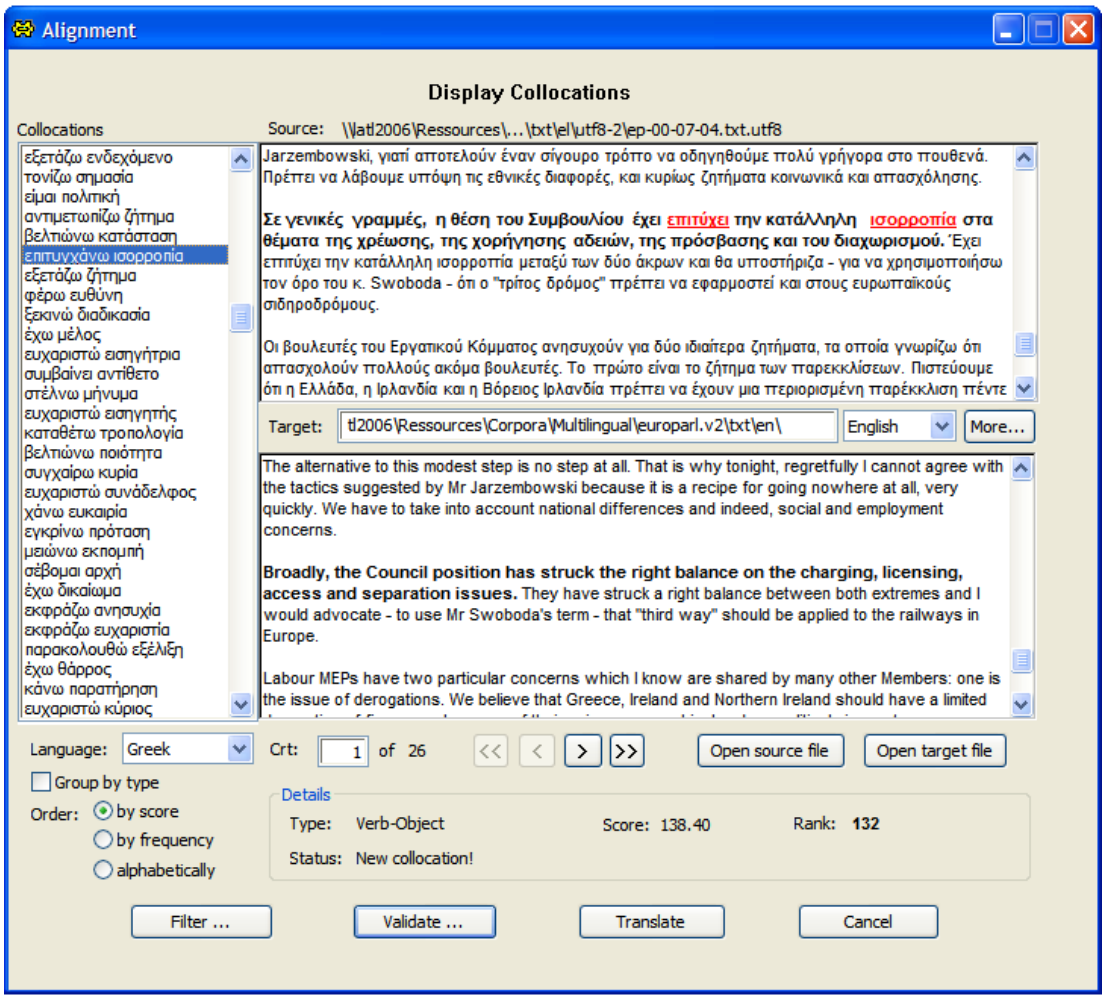

Figure 1: Screen capture of the parallel concordancer, showing an instance of the collocation επιτυγχάνω ισορροπία ('strike balance') and the aligned context in the target language, English.

The extractor is designed as a module which is plugged into the parser. After a sentence from the source corpus is parsed, the extractor traverses the output structure and identifies as a potential MWE the words found in one of the syntactic configurations listed in Section 3.

Once all MWE candidates are collected from the corpus, they are divided into subsets according to their syntactic configuration. Then, each subset undergo a statistical analysis process whose aim is to detect those candidates that are highly cohesive. A strong association between the items of a candidate indicates that this is likely to constitute a collocation. The strength of association can be measured with one of the numerous association measures implemented in our extractor. By default, the log-likelihood ratio measure (LLR) is proposed, since it was shown to be particularly suited to language data (Dunning, 1993).

In our extractor, the items of each candidate expression represent base word forms (lemmas) and they are considered in the canonical order implied by the given syntactic configuration (e.g., for a verb-object candidate, the object is postverbal in SVO languages like Greek). Even if the candidate occurs in corpus in a different morphosyntactic realizations, its various occurrences are successfully identified as instances of the same type thanks to the syntactic analysis performed with the parser.

4.2 Visualization

The extraction tool also provides visualization functions which facilitate the consultation and interpretation of results by users—e.g., lexicographers, terminologists, translators, language learners—by displaying them in the original context. The following functions are provided:

Filtering and sorting The results which will be displayed can be selected according to seve-

ral criteria: the syntactic configuration (i.e., users can select only one or several configurations they are interested in), the LLR score, the corpus frequency (users can specify the limits of the desired interval),[3] the words involved (users can look up MWEs containing specific words). Also, the selected results can be ordered by score or frequency, and users can filter them according to the rank obtained.

Concordance The (filtered) results are displayed on a concordancing interface, similar to the one shown in Figure 1. The list on the left shows the MWE candidates that were extracted. When an item of the list is selected, the text panel on the right displays the context of its first instance in the source document. The arrow buttons beneath allow users to navigate through all the instances of that candidate. The whole content of the source document is accessible, and it is automatically scrolled to the current instance; the component words and the sentence in which they occur are highlighted in different colors.

Alignment If parallel corpora are available, the results can be displayed in a sentence-aligned context. That is, the equivalent of the source sentence in the target document containing the translation of the source document is also automatically found, highlighted and displayed next to the original context (see Figure 1). Thus, users can see how a MWE has previously been translated in a given context.

Validation The tool also provides functionalities allowing users to create a database of manually validated MWEs from among the candidates displayed on the (parallel) concordancing interfaces. The database can store either monolingual or bilingual entries; most of the information associated to an entry—such as lexeme indexes, syntactic type, source sentence—is automatically filled-in by the system. For bilingual entries, a translation must be provided by the user, and this can be easily retrieved manually from the target sentence showed in the parallel concordancer (thus, for the collocation shown in Figure 1, the user can find the English equivalent *strike balance*).

[3]Thus, users can specify themselves a threshold (in other systems it is arbitrarily predefined).

5 Conclusion

We presented a MWE extractor with advanced concordancing functions, which can be used to semi-automatically build Greek monolingual/bilingual MWE lexicons. It relies on a deep syntactic approach, whose benefits are manifold: retrieval of grammatical results, interpretation of syntactic constituents in terms of arguments, disambiguation of lexemes with multiple readings, and grouping of all morphosyntactic variants of MWEs.

Our system is most similar to Termight (Dagan and Church, 1994) and TransSearch (Macklovitch et al., 2000). To our knowledge, it is the first of this type for Greek.

Acknowledgements

This work has been supported by the Swiss National Science Foundation (grant 100012-117944). The authors would like to thank Eric Wehrli for his support and useful comments.

References

Anna Anastasiadi-Symeonidi. 1986. *The neology in the Common Modern Greek*. Triandafyllidi's foundation, Thessaloniki. In Greek.

Ido Dagan and Kenneth Church. 1994. *Termight*: Identifying and translating technical terminology. In *Proceedings of ANLP*, pages 34–40, Stuttgart, Germany.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Elliott Macklovitch, Michel Simard, and Philippe Langlais. 2000. TransSearch: A free translation memory on the World Wide Web. In *Proceedings of LREC 2000*, pages 1201–1208, Athens, Greece.

Athina Michou. 2007. Analyse syntaxique et traitement automatique du syntagme nominal grec moderne. In *Proceedings of TALN 2007*, pages 203–212, Toulouse, France.

Angela Ralli. 2005. *Morphology*. Patakis, Athens. In Greek.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLING 2002*, pages 1–15, Mexico City.

Violeta Seretan. 2008. *Collocation extraction based on syntactic parsing*. Ph.D. thesis, University of Geneva.

Eric Wehrli. 2007. Fips, a "deep" linguistic multilingual parser. In *Proceedings of ACL 2007 Workshop on Deep Linguistic Processing*, pages 120–127, Prague, Czech Republic.