Converting computational dictionaries to robust local grammars for NLP within the Lexicon-Grammar methodological framework

Voula Giouli
Institute for Language and Speech Processing, Athena RIC
E-mail: voula@ilsp.athena-innovation.gr

Abstract

In this paper we present a suite of computational lexica that were developed in accordance with the Lexicon-Grammar formalism aiming at the formal description of the Greek language. We will elaborate further on the methodology adopted, the resulting encoding, and their conversion to local grammars.
Keywords: Lexicon-Grammar; lexicon-grammar tables; computational dictionaries; graphs; local grammars.

1. Introduction

Computational lexica, that is, dictionaries used by natural language processing systems, differ in many critical ways from Machine Readable dictionaries intended for human end-users. This paper describes work aimed at the development of computational lexica for the Greek language that encompass verb predicates denoting emotion and their encoding in a lexical resource on the basis of formal syntactic and semantic criteria. The focus will be on phenomena that need to be addressed during processing, and the overall contribution of the lexicographic data to the symbolic processing of text and in view of specific applications (Information Extraction, sentiment analysis, etc).

The paper is outlined as follows: the theoretical background that forms the basis of the intended dictionaries is presented in section (2); the methodological principles of entry selection and encoding are outlined in section (3), whereas the LG Tables developed and their conversion to local grammars are presented in sections (4) and (5) respectively. Finally, conclusions and future research are reported in section (6).

2. Background

The Lexical Resources hereby presented were constructed in accordance with the Lexicon-Grammar (LG) methodological framework (Gross 1975), (Gross 1981). Being initially conceived of as a model of syntax, the formalism is based on the principles of Transformational Grammar (Harris, 1951; 1964; 1968). In this respect, the point of departure is the principle that the *unit of meaning* is *located at the sentence rather than the word level*. To this end, the elementary sentence is defined as having the form *Subject – Verb – Object*,

and linguistic analysis consists in converting each elementary sentence to its predicate-argument structure. Additionally, main complements (subject, object) are separated from other complements (adjuncts) on the basis of formal criteria; adverbial complements (i.e., prepositional phrases) are considered as crucial arguments only in the case that they characterize certain verb frames:

(1) John removed the cups from the table.

To cater for a more fine-grained classification, and the creation of homogenous word classes, this formal syntactic definition is further coupled with *distributional properties* associated with words, i.e., types of prepositions, features attached to nouns in subject and complement positions, etc.

Moreover, transformation rules, construed as equivalence relations between sentences, further generate *equivalent structures*, which are also encoded as properties where appropriate.

Each table is defined by a set of distinct properties (syntactic, distributional, and semantic) and includes all lexical items sharing these properties. In this sense, tables are homogenous in that they comprise predicates that share the same properties. Lexical units with more than one usage or meaning, have been treated as separate lexical items (possibly represented in different tables) and syntactic and semantic properties are assigned to each meaning thereof. A set of features that are appropriate for the description of a syntactic category concerning grammatical (i.e., past participle form) or syntactic information (i.e., passive transformation, causative-inchoative alternation, etc.) and lexical choice (i.e., preposition a given predicate is subcategorized for) is applied to all entries, and their linguistic validity was checked on the basis of corpus evidence. At the intersection of a row corresponding to a lexical item and a column corresponding to a feature, the cell is set to '+' if it is valid or '-' if is not.

It becomes evident, therefore, that the resulting resources are rich in *linguistic information* (syntactic structure, distributional properties and permitted transformational rules), which is encoded *formally* in the so-called LG tables.

In the next sections, we will elaborate further on the development of computational lexica for the Greek language and their conversion to local grammars that are applicable for Natural Language Processing tasks (i.e., parsing, Sentiment Analysis, Machine Translation, etc). The computational dictionaries comprise verbal predicates denoting emotion or emotional state in Greek.

3. Methodological Principles

The present work was performed on the basis of corpus evidence. To this end, we made extensive use of the Hellenic National Corpus (HNC), a large reference corpus for the Greek language (Hatzigeorgiou et al, 2000). Additionally, a suite of specialized corpora that were developed to guide sentiment studies in multimodal (Mouka et al., 2012) and in textual (Giouli and Fotopoulou, 2013) data was used. Thus, the resulting Greek Sentiment Corpus, that amounts to c. ~250K tokens, comprises audiovisual material (movies dialogues), and texts selected manually from various sources over the web. The resulting corpus was subsequently employed to (a) populate the sentiment lexicon under construction; (b) identify the properties of the selected predicates, and (c) to guide and test the local grammars developed.

A core lexicon of emotion verbs extracted from existing lexical resources and corpora has been manually updated and extended; the semantic class of verb predicates denoting emotion was then defined based on a set of formal criteria. The verbs were then classified on the basis of their syntactic and semantic properties.

Verbs with more than one usage/meaning have been treated as separate lexical items, and their properties are assigned to each one thereof.

The set of properties identified to hold for the semantic class at hand has been applied to all verb predicates, and their linguistic validity was checked against corpus evidence.

3.1. Defining the semantic class of emotions: verb selection

Candidate verbs were manually selected from a concept-based Lexicon for Modern Greek, namely, *Antilexicon* (Vostantzoglou, 1962). A set of appropriate lexical semantic tests were then employed as a formal device guiding the selection of a core vocabulary of emotions that covers the grammatical category of verbs, and the delineation of the semantic class of emotions, leaving semantically and conceptually related verbs aside for future treatment.

Consequently, the so-produced list of verb predicates was further enriched with entries located in existing standard lexicographic resources for Greek via synonyms identification. Additionally, specialized electronic resources for the English language were also consulted, namely FrameNet (Fillmore, 2001) and WordNet (Fellbaum, 1998), and Greek data were supplemented by obtaining translations of the English entries. This process resulted in a list of 339 verbs predicates that were finally selected for further study.

4. The Lexicon – Grammar of verb predicates denoting emotion

LG tables describe formally from a linguistic point of view the argument structure, distributional properties and possible transformations of the included predicates. Each class is represented by a table that includes all lexical items of that class.

Classification of the predicates at hand was performed on the basis of the following axes: (i) syntactic information (i.e, subcategorisation information); (ii) selectional restrictions (+Hum/ -Hum) imposed over their Subject and Object complements; and (iii) transformation rules. The predicates at hand were classified in 5 classes each one depicted in a separate LG table. More precisely, as far as syntactic structure is concerned, the predicates under consideration were identified to appear in both transitive and intransitive constructions being represented as *N0 V N1* and *N0 V* respectively. Certain verbs also allow for a prepositional phrase complement represented as *N0 V Prep N1*[1] configurations. A close inspection over the data revealed the relationship between the N0 or N1 complements that denote the Experiencer of the emotion (i.e., the entity feeling the emotion). In two of the resulting classes the Experiencer is projected as the structural Subject of the verb, whereas the Theme or Stimulus is projected as their structural object. Similarly, the remaining 3 classes realize the Theme/Stimulus as the subject and the Experiencer as their object, their distinguishing property being their participation in unaccusative and middle constructions, the latter being linked to the implicit presence of an Agent (middle) and the absence of an Agent (unaccusative). These properties have been checked for the whole range of lexical data based on both linguistic introspection and corpus evidence.

A number of Harrisian constructions and transformations (Harris, 1951; 1964; 1968) have been extensively utilized within the LG formalism to define syntactically related and semantically equivalent structures. Apart from passivisation and middle alternation constructions - also relevant to emotion predicates - the *restructuring* transformation has been accounted for (Guillet and Leclère, 1981):

(2) Ο Γιάννης θαυμάζει τη Μαρία για το θάρρος της.

O Jianis thavmazi ti Maria jia to tharos tis

The John admires the Maria for the courage-her.

John admires Maria for her courage.

(3) Ο Γιάννης θαυμάζει το θάρρος της Μαρίας.

[1] Adopting the LG notation, N0 denotes a Noun in *Subject* position of a given verb V, whereas, N1 denotes its *Object*.

O Jianis thavmazi to tharos tis Marias

The John admires the courage the.SG.GEN MariaSG.GEN

John admires Maria’s courage.

Moreover, each verbal predicate was also coupled with morphologically-related adjectives and nouns, and the alignment of semantically equivalent nominal, verbal and adjectival structures was performed thereof. A number of semantically equivalent paraphrases of the verbs with the morphologically related nouns and adjectives were also encoded in the tables. Finally, following common lexicographic practices, each lexicon entry is coupled with an example. A sample of a LG table is given in Figure 1 below.

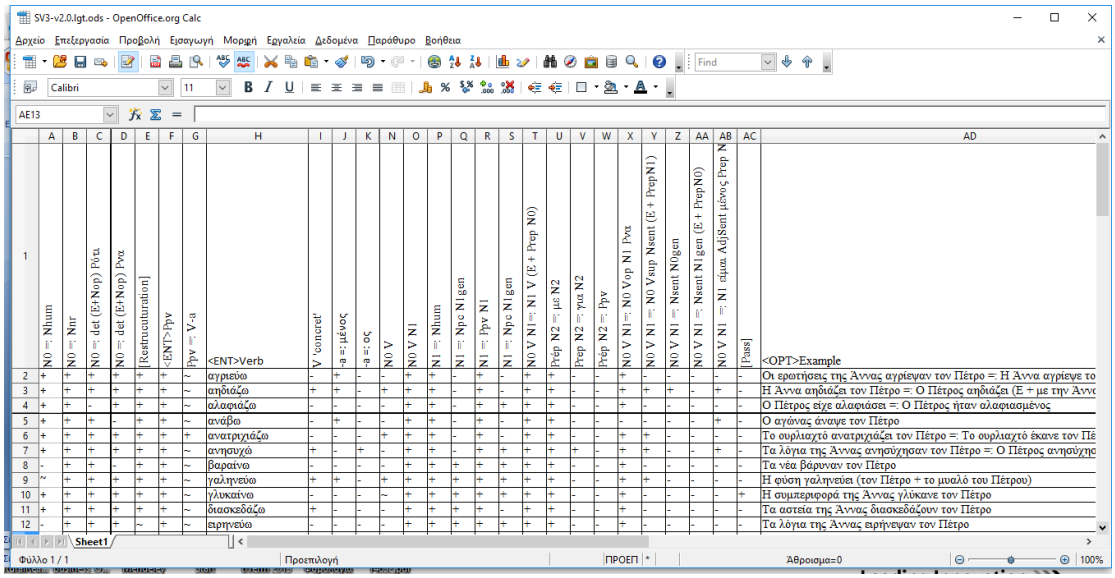

Figure 1. LG table

5. Transforming Lexicon-Grammar tables to local grammars

Being initially developed to serve as a means of linguistic description, this framework has, never-the-less, been proved to be applicable for the construction of robust computational lexica. And although it has been claimed (Mathieu and Tolone, 2008) that the information is not directly exploitable for NLP applications due to the fact that certain pieces of information are not formally encoded or are implicit, a number of works (Hathout and Namer 1998, Danlos and Sagot 2009) have successfully managed to reformat LG tables in efficient large-scale NLP lexica. To this end, we have tried to exploit information available in the tables and make the mappings that are necessary for Natural Language Processing tasks, the focus being on parsing. To this end, a library of local grammars for emotion predicates has been constructed

modeling structures in the corpus.

Local grammars (also referred to in the literature as graphs) are algebraic grammars formulated as combinations of sequences of grammatical symbols in the form of regular expressions that describe natural language. In this sense, they are a powerful tool to represent the majority of linguistic phenomena in an intuitive manner.

We made use of the UNITEX platform (Paumier 2013) for creating the graphs and then compiling them into finite state transducers. UNITEX consists of three modules, namely, corpus handling, lexicon development and grammar development that are integrated into a single intuitive graphical user interface.

Based on the Lexicon-Grammar tables developed for the verbal predicates (c.f. section 4 above), we initially created five parameterized graphs manually; these graphs depict the syntactic and semantic properties that are depicted in each one of the 5 LG tables constructed.

At the next stage, using the built-in functionality of UNITEX, we constructed automatically a set of graphs on the basis of the information provided in the LG tables for each one of the lexical entries and the absence or presence of a given property. In this regard, each graph represents the syntactic and semantic properties of a given predicate. The outcome of this procedure is 339 graphs.

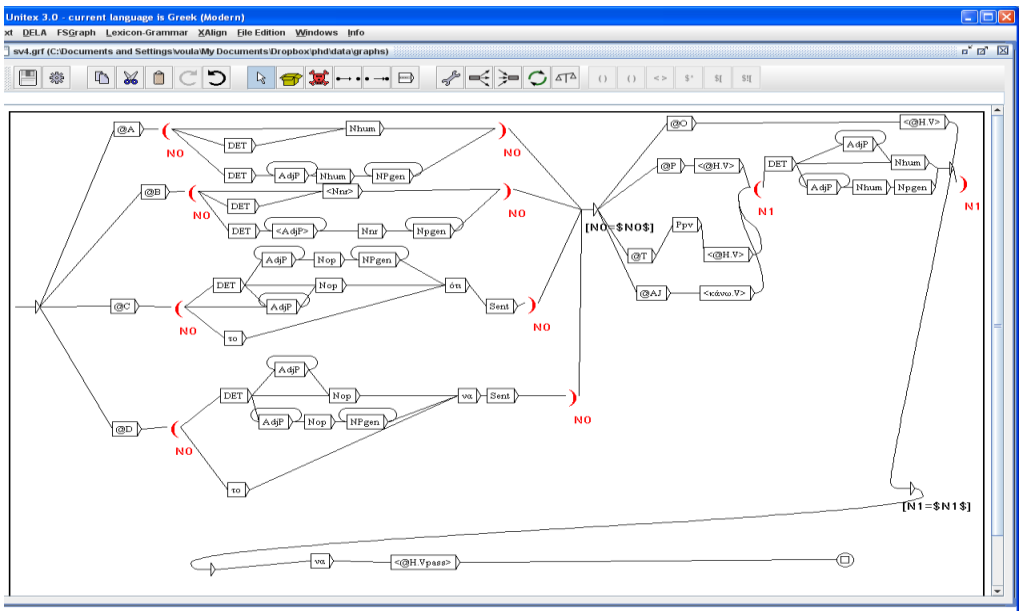

Figure 2. Developing a local grammar

Moreover, they are compiled into finite state transducers that transform input text by inserting or removing special markers, as shown in Figure (2).

6. Conclusions

We have presented a set of computational dictionaries in the form of LG tables that depict the syntactic and semantic properties of emotion predicates in Greek and their conversion into local grammars. The suite of resources also comprises LG tables for nominal predicates along with verbal multi-word expressions denoting emotion. Our efforts have been oriented towards developing a rule-based system that will eventually recognise emotion expressions in texts and the participants in the emotion event. Future work has been planned towards the exploitation of other properties that are encoded in the LG tables, as for example the restructuring property as a facet of the aspect-based sentiment analysis and the conversion of the enriched LG tables to a standardised lexical format. Finally, the validation of the final resource is due against the manually annotated corpus.

7. References

Antilexicon i Onomastikon tis Neas Ellinikis Glossis. Athens. Vostantzoglou, Th. 1962.

Danlos, L., Sagot, B. (2009). Constructions pronominales dans Dicovalence et le lexique-grammaire: Integration dans le Lefff. *Actes du 27e Colloque international sur le lexique et la grammair*e.

Fellbaum, C. (1998). (ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fillmore, C., Wooters, C., Baker, K. (2001). Building a Large Lexical Databank Which Provides Deep Semantics. In *Proceedings of the Pacific Asian Conference on Language, Information and Computation*. Hong Kong.

Giouli, V. and A. Fotopoulou (2013). Developing Language Resources for Sentiment Analysis in Greek. In *Proceedings of the International Conference in Greek Linguistics*. ICGL, Rhodes, Greece.

Giouli, V., and A. Fotopoulou (2012). Emotion verbs in Greek. From Lexicon-Grammar tables to multi-purpose syntactic and semantic lexica. In *Proceedings of the XV Euralex International Congress* (EURALEX 2012). August 2012, Oslo, Norway, pp. 485-492.

Giouli, V., and A. Fotopoulou (2012). Towards developing Local Grammars of verb predicates denoting Emotion in Greek: a Lexicon-Grammar account. In *Proceedings of the 31st International Conference on Lexis and Grammar 2012*. Nové Hrady, Czech Republic 2012, September 19-22, 2012.

Gross, M. (1975). *Méthodes en syntaxe*. Régime des constructions complétives. Hermann, Paris.

Gross, M. (1986). Lexicon-grammar: The Representation of Compound Words. In Proceedings of the 11th Coference on Computational Linguistics, pp. 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Harris, Z.S. (1951). *Methods in Structural Linguistics*. The University of Chicago Press, Chicago.

Harris, Z.S. (1964). *The Elementary Transformations*. In T.D.A.P. University of Pennsylvania 54, Pennsylvania.

Harris, Z.S. (1968). *Mathematical Structures of Language*. Wiley, New York.

Hathout, N.,and F. Namer (1998). Automatic Construction and Validation of French Large Lexical Resources: Reuse of Verb Theoretical Linguistic Descriptions. In *Proceedings of the 1st Language Resources and Evaluation Conference*, Granada, Spain.

Hatzigeorgiu, N., M. Gavrilidou, S. Piperidis, G. Carayannis, A. Papakostopoulou, A. Spiliotopoulou, A. Vacalopoulou, P. Labropoulou, E. Mantzari, H. Papageorgiou, and I. Demiros (2000) Design and Implementation of the Online ILSP Greek Corpus. In *Proceedings of the 2nd Language Resources and Evaluation Conference* (LREC, 2000), Athens, Greece.

Matthieu, C. and Tolone, E. (2008). A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. *Actes du 27e Colloque international sur le lexique et la grammaire*, L'Aquila, 10-13 Septembre 2008.

Mouka, E., V. Giouli, A. Fotopoulou and I.E. Saridakis (2012). Opinion and emotion in movies: a modular perspective to annotation. *4th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals* (ES³ 2012), Istanbul, Turkey, 26 May 2012.

Paumier, S. (2003). UNITEX User Manual.