

Review of Parsing in Modern Greek - A New Approach

George Chorozoglou

*School of Engineering/Department of Informatics and Computer Engineering
University of West Attica
Egaleo, 122 43, Greece*

gchorozoglou@uniwa.gr

Nikolaos Z. Zacharis

*School of Engineering/Department of Informatics and Computer Engineering
University of West Attica
Egaleo, 122 43, Greece*

nzach@uniwa.gr

Evangelos C. Papakitsos

*School of Engineering/Department of Industrial Design and Production Engineering
University of West Attica
Egaleo, 122 41, Greece*

papakitsev@uniwa.gr

Eleni Galiotou

*School of Engineering/Department of Informatics and Computer Engineering
University of West Attica
Egaleo, 122 43, Greece*

egali@uniwa.gr

Apostolos Giovanis

*School of Administrative Economics and Social Sciences/Department of Business Administration
University of West Attica
Egaleo, 122 44, Greece*

agiovanis@uniwa.gr

Abstract

This paper deals with the attempts of parsing Modern Greek language and the techniques used. Most attempts are described without being thoroughly analyzed for the purpose of reviewing and comparing methods. The desire to use Link and Relational Grammars for the syntactic analysis of Greek language is also stated. Link Grammar has a robust algorithmic structure, while Relational Grammar present results in a standard and commonly understandable manner. Therefore, their combination is expected to facilitate the presentation of results in a more user-friendly manner, and to exhibit a better processing efficiency.

Keywords: Natural Language Processing, Greek Language, Syntactic Analysis, Link Grammar, Relational Grammar.

1. INTRODUCTION

The Greek language belongs to the group of Indo-European languages and has some special features in relation to the other languages. It is a synthetic, inflectional or fusional language [1], [2]. Its analysis, like any language during processing (Natural Language Processing) can be done on a phonological, morphological, syntactic, semantic and pragmatic level. The herein study is focused on examining the syntactic level, as being assisted by the morphological level.

This article starts with reference to issues of syntactic analysis, moves on to the efforts for the syntactic analysis of the Greek language and ends with the new approach, i.e., the usage of the combination of Link [3], [4] and Relational Grammar [5], [6], [7] for this purpose.

2. SYNTACTIC ANALYSIS

Syntactic analysis is a multilevel process. It aims to identify each constituent word(s) and assign syntactic roles (e.g., subject, object). In addition, the creation of a syntax tree clearly shows the dependencies between all the constituents (terms) of each sentence. Thus *swallow* or *partial* parsing refers to a simple analysis, while *deep* or *full* parsing presents us with a complete syntactic tree as a result [8], [9].

One of the basic assumptions of modern syntactic theory is that language sentences are not simply considered as a sequence of elements, but as hierarchically arranged constituent structures. The permissible structures are described by grammar, which is a set of phrase structure rules that define how sentences are structured by phrasal categories and then how phrasal categories are structured by lexical categories [10].

On the subject of analysis process approach, it should be mentioned that there are two basic methods. The first is rule-based, following a grammatical formalism, and the second is statistical, which is based on statistically derived rules [11], [12].

The Greek language is distinguished for its rich morphology and the relatively free constituent order of each clause. The many inflectional types in addition with the various cases that are found in nouns and adjectives help in recognizing the various syntactic terms, such as, e.g., the subject. On the other hand, free order can make it difficult to recognize other syntactic terms. The result is ambiguity, i.e., the application of alternative syntactic rules to a sentence, with the production of different analyses. In particular, computational ambiguity can be lexical (ambiguous in terms of grammatical recognition), semantic (give different meanings) and/or syntactic (ambiguous in terms of syntactic analysis). A good example is the phrase:

το αγόρι βλέπει το κορίτσι
(the boy sees the girl)

where the object cannot be distinguished from the subject, since the words *αγόρι* (boy) and *κορίτσι* (girl) are of the neuter gender, where accusative and nominative cases have the same type. Unlike other languages, the Greek language requires specific cases on subject and object. Ambiguity resolution can be removed in oral speech by emphasizing the corresponding term, while computationally by utilizing the relevant order of sentence terms.

3. EFFORTS FOR THE SYNTACTIC PARSING OF GREEK

The herein research methodology initially followed a bottom-up approach (inductive reasoning), by collecting and studying all the accessible attempts of developing syntactic parsers for Greek (there are also attempts either not made public or not accessible). The features that have been analyzed included their grammar formalisms, parsing algorithms and performance, with the purpose of discovering potential deficiencies or limited coverage of syntactic phenomena. The scope of this research is to propose a new design for parsing Greek language that will be algorithmically robust and user-friendly in presenting the results to non-expert in linguistics, for facilitating its application in various fields.

In one of the first attempts on the processing of Greek language (1995) [13], [14], the basic principles of the theory of HPSG (Head-driven Phrase Structure Grammar) were adopted [15]. The position of the constituents also does not play a special role in the formulations of HPSG, which gives great emphasis to the detailed description of the grammatical features, as in the Greek language. The formulation of linguistic knowledge is done by using hierarchical networks. The description of the morphological properties of the words and the syntactic behavior of the components of the sentences is done by the formalization of the frames and the development of a rectangular inheritance network, i.e., some features to be inherited from one frame and some from another (by using object-oriented techniques in Pascal programming language). The results of the morphological dictionary and the declension rules are used by the syntactic analyzer to

determine the syntactic structure. DINUS (dis = dual + nous = mind), as the parser is called, was developed and based on the operating principles of a Shift-Reduce parser. Decisions are made on the basis of linguistic knowledge and experience from exemplary sentence analyses (recorded correspondences from a computer training process). To limit the alternative analyses, a "plausibility" index, created using situations that have been dealt with in the past, is used.

In another attempt (1997) [16], [17], a syntactic parser based on the PATR II formalism [18] was implemented, which incorporates features of a unification-based grammar. The PC-PATR environment was used for this application, i.e., a version of PATR for PCs with an integrated bottom-up chart parser. A grammar rule file and a dictionary file are required to describe grammar with PC-PATR. Regarding the morphological parser, the Two-Level Morphology Model, developed by Kimmo Koskienniemi, was used [19]. The two levels concern the level of lexical representation, i.e., the word as a series of morphemes without any alteration, and the one of surface representation, which is the orthographically correct form as it has resulted from the application of all phonological alterations. For its implementation, a special programming environment was used, called PC-KIMMO (named after *Kimmo* Koskenniemi) and created by Lauri Karttunen. PC-KIMMO must have a dictionary and two-level rules as separate files, represented as Finite State Automata (FSA) for the dictionary and Finite State Transducers (FST) for the rules. In a separate file, the rules concerning the grammatical features of the morphemes are also stated. PC-PATR works directly with the PC-KIMMO environment.

There were also developed:

- a module for grapheme-to-phoneme conversion and vice-versa (*BD-GRAPH*), based on FST,
- two original word-entry algorithms in Directed Acyclic Word Graphs (*DAWGs*) - one composes deterministic *DAWGs* and the other non-deterministic ones,
- an independent morphological processor (*POLYMORPH*) and
- a temporal semantic parser (*TEMPO*) for the detection and analysis of the quantifiers of time that can be used either independently or in collaboration with other semantic units.

In a different approach based on empirical methods (1997) [20], Greek is characterized as a language with Quasi Free Word Order and the syntactic text processor is based on a model that is not bound by a known modern grammar formalism but a reliable and efficient pattern matching algorithm with constraints. Empirical methods are based on observation, experience and experimentation. Through the processing of a large number of texts, problems are recorded and analyzed, a statistical approach is made and heuristic rules are applied for the effective handling of ambiguities and unusual grammatical syntaxes. The parser divides the periods into sentences and then identifies these sentences and their components. This is done using a morphological dictionary (lexical database), some predefined keywords and the application of two sets of heuristic rules: one for the split of periods and one for the recognition of sentences (main, subordinate, parenthetical but also judgment, wish, exclamatory, interrogative). The keywords and the heuristic rules were found after a study and statistical analysis of the Greek language. Especially heuristic rules are called in order of priority to avoid unwanted confusion. To identify the type of phrases (noun, verb, adjective, prepositional, and adverbial), a pattern matching algorithm, depicting word sets (e.g., article + noun), is used. In addition to the syntactic parser, a three-level stylistic parser has been created that classifies texts, as well as a terminology interpreter that handles issues related to the ordering of terms and their interpretation. Everything has been implemented in the Prolog programming language.

The next attempt (2000) concerns partial (surface) free text syntactic parsing [21], using Finite State Automata techniques and sub-categorization frameworks. Non-recursive rules for syntactic analysis take the form of Regular Expressions, which translate into FSA or FST. The analysis is deterministic and no backtracking is done. Phrases and sentences are firstly identified by FST and then the identified components are labeled through a sub-categorization dictionary and a

pattern matching mechanism. Ambiguities are either not resolved or decisions are made that favor one analysis (more likely to be correct) than the rest.

Continuing this search, the use of Unification Grammar can be seen for the theoretical description of the Greek language and HPSG for the syntax of noun phrases, through an effort (2005) [22] with main theme the automatic retrieval of syntactic information (recognition of complements and retrieval of sub-categorization frames of verbs) in text/corpora using machine learning techniques. The PATR-II formalism with the PC-PATR environment is used to implement the former. The dictionary, which has about 20,000 entries with morphological information, was created mainly automatically by editing words with text/corpora, while the grammar file contains grammar rules in combination with the feature restrictions, corresponding to each rule. A problem for the above implementation is the inability to attach prepositional phrases with the number of syntactic analyses, increasing as the words increase. Regarding HPSG [23], the Attribute Logic Engine (ALE) [24] was used as the development environment with a built-in bottom-up chart parser. Lexical entries provide, in addition to morphological information, a very important part of syntactic and semantic information, while the more general syntax rules limit erroneous analysis results. Of course, HPSG has a complexity in its description.

FipsGreek syntactic parser [25] was developed at the University of Geneva and is part of the multilingual Fips syntactic analysis model [26]. It is based mainly on the Generative Grammar [10], which refers to general rules for the languages used (French, German, Italian, English & Spanish) and additional rules to cover the specifics of the Greek language. It is a syntactic parser of deep structure and is based on the mechanisms of *projection*, *merge* and *move*. There is a morphological dictionary as a lexical database, which contains morphological and semantic information along with grammar, consisting of rules and procedures to combine lexical elements with each other. Unfortunately, the relatively low results of the Greek parser in relation to the other languages are due to the fact that the program is still under development and the grammar of Greek language has not been completed.

Mnemosyne (2015), which is a complete natural language processing system, containing a syntactic parser, was used to construct a Greek grammar checker [27], [28], [29]. This software is implemented with Java. Its formalism abides by the logic of unifying grammars, it is called Kanon, and includes rules that apply to both context-free grammar and context-sensitive grammar.

Another remarkable project is the integration of Greek Language repository in spaCy (<https://spacy.io/models/el>). SpaCy is an open source library of Natural Language Processing (NLP) programs written in Python & Cython. Unlike NLTK, which was developed more for educational purposes, spaCy (<https://spacy.io/usage>) is designed to build NLP applications. There are three predefined statistical models for the Greek language, with different capabilities per model. It has capabilities for *text tokenization*, *sentence splitter*, *lemmatizer*, *Part of Speech Tagger*, *Named Entity Recognition*, recognizing lexical attributes such as numbers, urls & emails, *dependency tagging (DEP Tagger)* and detecting similarities between texts. The project "Google Summer of Code 2018" (<https://github.com/eellak/gsoc2018-spacy>) describes many of the above features.

For the completeness of this article, two attempts should be referred to that have been made by creating parsers that do not have a complete grammar, but tools for syntactic analysis. The first one is called "Text-Unifier" (Κείμενοποιοητής) [30] (1995) and "is a means of holistic representation of linguistic knowledge, regardless of level". Any text could have been edited as long as it was given the appropriate grammar that had to abide by the logic of the Unifying Grammars. It was based on the PATR II formalism and was written in C. The second one (<http://ioperm.org/lfg-parser.html>) (2014) is based on the Lexical Functional Grammar (LFG) formalism written in Java and is a Master thesis in the framework of the Postgraduate Program "Technoglossia" [31].

4. A NEW APPROACH: USING LINK AND RELATIONAL GRAMMAR

For the syntactic analysis of the Greek language, two other grammars will be evaluated that have not been used so far: Link Grammar (LG) in conjunction with Relational Grammar (RG). The first offers a robust algorithmic structure while the second delivers the results in a standard and commonly understandable manner. The use of these two grammars is designed for the first time in Greek and is an original attempt for achieving better results. The analysis with the transformative grammars concerns the relations of words throughout the text, while in the Link Grammar it has to do with the connections in neighboring words.

Link Grammar was developed by Sleator & Temperly (1995) [3], [4], with their original and classic articles to determine the limits of grammar and inspire later works. It includes a set of words (terminals) contained in a dictionary. A sequence of words belongs to this grammar (i.e., produced by it) if there are links (arcs) that connect them under the following conditions:

- Links should not intersect (Planarity).
- Links connect all words in the sequence (Connectivity).
- Links meet the linking requirements for each word (Satisfaction).

The linking requirements for each word are registered in the dictionary. It should be noted that they do not form syntactic trees, since some words are connected only to their left or only to their right or left and right at the same time. A simple representation can be seen in Figure 1. It has been applied to English, Russian, Persian, Arabic, German, Hebrew, Lithuanian, Vietnamese, Kazakh and Turkish.

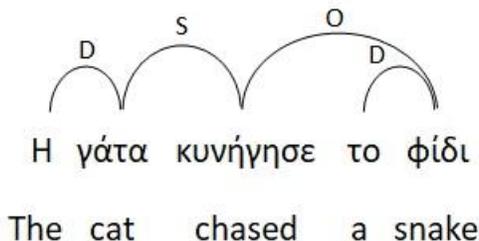


FIGURE 1: Simplified sentence form in LG where D - Determiner, S - Subject and O - Object.

On the other hand, Relational Grammar was developed by Perlmutter and Paul Postal in the early 1970s [5], [6], [7]. In this theory, the set of relationships recognized includes the subject, the direct and the indirect object. There is a hierarchy, where the subject is marked by {1}, the direct object by {2} and the indirect object by {3}. A generic representation of Relational Grammar can be seen in Table 1.

1	V(erb)	2	3
Ο Νίκος <i>Nick</i>	έδωσε <i>gave</i>	ένα φιλί <i>a kiss</i>	στην Ευρυδίκη <i>to Eurydice</i>

TABLE 1: Representation of Relational Grammar.

Based on these two grammars, the Greek language will be represented/described and a syntactic parser will be constructed that can be used in various applications. Such an application that is planned to be initially developed concerns the Sentiment analysis regarding the opinion mining of clients for organizations that provide tourism services; a matter of great concern for the local economy. Similar attempts of Sentiment analysis have been made for other sectors of the economy [32] and of opinion mining in other languages than English [33], as in this case (Greek). In addition, a platform of syntactic analysis will be ready for supporting various NLP applications, like machine translation or information retrieval.

5. CONCLUSIONS

The aim of this review was to collect and evaluate the existing parsers for the Greek language. After enquiring all the accessible efforts that have been made in the construction of syntactic analyzers for Greek, the methods that have been used, the grammars that have been based on, but also the tools that have been used, several points of improvement have been detected. Accordingly, a first attempt is made herein to give some basic elements for creating a new syntactic analyzer (parser), which will be based on Link Grammar and Relational Grammar, for improving the sentence processing and the presentation of parsing results. After the implementation of the parser and the tests of the above combination of grammars in a linguistic environment that concerns the field of texts from tourism organizations, it is aimed to extract (Information Extraction) and retrieve Information (Information Retrieval) that will help to evaluate the services that these organizations provide. The initial interest of application is focused on Sentiment analysis with the final product to be a complete software package.

6. REFERENCES

- [1] A. Ralli. Compounding in Modern Greek, *Studies in Morphology* vol. 2. Springer, 2013.
- [2] C. Charitonidis. "Polysynthetic Tendencies in Modern Greek". *Linguistik Online*, vol. 34(2), pp. 17-40, 2008.
- [3] D. Sleator and D. Temperley. "Parsing English with a Link Grammar" in *Proceedings of the Third International Workshop on Parsing Technologies*, Tilburg, Netherlands and Durbuy, Belgium, 1993, pp. 277-292.
- [4] D. Sleator and D. Temperley. "Parsing English with a Link Grammar". Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.
- [5] B. J. Blake. *Relational grammar*. London: Routledge, 1990.
- [6] D. M. Perlmutter (editor). *Studies in Relational Grammar 1*. Chicago: University of Chicago Press, 1983.
- [7] D. M. Perlmutter and C. G. Rosen (editors). *Studies in Relational Grammar 2*. Chicago: The University of Chicago Press, 1984.
- [8] N. Indurkha and F. J. Damerau (editors). *Handbook of Natural Language Processing 2nd Edition*. Chapman & Hall/CRC, New York, 2010.
- [9] D. Jurafsky and J. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd Edition)*. Prentice-Hall, New Jersey, 2009.
- [10] N. Chomsky. *The Minimalist Program*. Cambridge, MA: MIT Press, 1995.
- [11] S. Abney. "Statistical methods in language processing". *WIREs Cognitive Science*, vol. 2, pp. 315-322, 2011.
- [12] J. Nivre. "On Statistical Methods in Natural Language Processing". *Promote IT: Second Conference for the Promotion of Research in IT at New Universities and University Colleges in Sweden*, pp. 684-694, 2002.
- [13] A. Draggiotis. "Contribution to the Development of Natural Language Parsing Methods Referring especially to Sentences in Modern Greek." (in Greek) Ph.D. Dissertation, National and Kapodistrian University of Athens, Greece, 1995.
- [14] A. Draggiotis, M. Grigoriadou, and G. Philokyprou. "The DINOUS parser", *Natural Language Engineering*, vol. 4(2), pp. 145-173, 1998.

- [15] C. Pollard and I. Sag. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press, 1992.
- [16] K. Sgarbas, N. Fakotakis and G. Kokkinakis. "A PC-KIMMO-Based Morphological Description of Modern Greek". *Literary and Linguistic Computing*, vol. 10, Issue 3, pp. 189-201, 1995.
- [17] K. Sgarbas. "Techniques for Automatic Morphosyntactic Analysis of Modern Greek Language." (in Greek) Ph.D. Dissertation, Department of Electrical and Computer Engineering, University of Patras, Greece, 1997.
- [18] S. M. Shieber. "The Design of a Computer Language for Linguistic Information" in *Proceedings of Coling '84, 10th International Conference on Computational Linguistics*, Stanford University, California, 2-7 July, 1984, pp.362-366.
- [19] K. Koskenniemi. "Two-level Morphology: A General Computational Model for Word-Form Recognition and Production". Publication No.11, University of Helsinki, Department of General Linguistics, 1983.
- [20] S. Michos. "Computational Syntactic Processing of Modern Greek Texts using Empirical Methods." (in Greek) Ph.D. Dissertation, University of Patras, Greece, 1997.
- [21] S. Boutsis, P. Prokopidis, V. Giouli and S. Piperidis. "A robust parser for unrestricted Greek text" in *Proceedings of the 2nd Language Resources and Evaluation Conference*, Athens, 2000, pp. 467-474.
- [22] K. L. Keramidis. "Learning of syntactic dependencies and development of Modern Greek grammars." (in Greek) Ph.D. Dissertation, University of Patras, Greece, 2005.
- [23] V. Kordoni and N. Julia. "Building multi-purpose linguistic resources for Modern Greek: a deep Modern Greek Grammar". *Workshop on Balkan Language Resources and Tools*, Thessaloniki, Greece, 2003.
- [24] B. Carpenter and G. Penn. *ALE. The Attribute Logic Engine. Users Guide. Version 3.2.1*, 2001.
- [25] A. Michou. "FIPSGREEK: A parsing model for Greek language". *8th International Conference on Greek Linguistics*, pp. 999-1012, 2007.
- [26] E. Wehrli. "Fips, a "Deep" Linguistic Multilingual Parser" in *DeepLP '07: Proceedings of the Workshop on Deep Linguistic Processing*, June 2007, pp. 120-127.
- [27] P. Gakis. "Design, construction and evaluation of the Greek grammar checker." (in Greek) Ph.D. Dissertation, University of Patras, Greece, 2015.
- [28] P. Gakis, C. Panagiotakopoulos, K. Sgarbas, C. Tsalidis and V. Verykios. "Construction of a Modern Greek Grammar Checker Through Mnemosyne Formalism" in: A. Ronzhin, R. Potapova, N. Fakotakis (eds) *Speech and Computer. SPECOM 2015. Lecture Notes in Computer Science*, vol. 9319. Springer, 2015.
- [29] P. Gakis, C. Panagiotakopoulos, K. Sgarbas, C. Tsalidis and V. Verykios. "Design and construction of the Greek grammar checker". *Digital Scholarship in the Humanities*, vol. 32(3), pp. 554-576, 2016.
- [30] Y. Maistros. "Greek Language Processing." (in Greek) Ph.D. Dissertation, National Technical University of Athens, Greece, 1996.
- [31] P. Minos. "Development of a natural language syntactic analyzer using LFG formalism." (in

Greek) Master thesis, National & Kapodistrian University of Athens and National Technical University of Athens, Greece, 2014.

- [32] A. Liapakis, T. Tsiligiridis, C. Yialouris and M. Maliappis. "A Corpus Driven, Aspect-based Sentiment Analysis To Evaluate In Almost Real-time, A Large Volume of Online Food & Beverage Reviews". *International Journal of Computational Linguistics (IJCL)*, vol. 11(2), pp. 49-60, 2020.
- [33] N. T. Mhaske and A. S. Patil. "Opinion Mining Techniques for Non-English Languages: An Overview". *International Journal of Computational Linguistics (IJCL)*, vol. 7(2), pp. 55-64, 2016.