DELOS: An Automatically Tagged Economic Corpus for Modern Greek

Katia Lida Kermanidis, Nikos Fakotakis, George Kokkinakis

Wire Communications Laboratory
Department of Electrical and Computer Engineering
University of Patras, 26500 Rio, Greece
{kerman, fakotaki, gkokkin}@wcl.ee.upatras.gr

Abstract

Text corpora resources have become an essential tool for Natural Language Processing tasks over the past years. A wide range of applications like information retrieval, ontology and terminology extraction require a sufficiently large corpus but of restricted domain. Manual tagging of such a corpus is very costly, making automatic annotation by a set of linguistic tools a very challenging idea. DELOS, described in this paper, is a Modern Greek corpus of economic domain consisting of 5 million word tokens, which is automatically tagged for morphology and shallow syntactic relations. The annotating tools described are embodied in an integrated system and their application to the corpus is performed using the GATE text engineering platform. The system output is a textual database marked up with the annotation tagset in plain text as well as in XML format.

1. Introduction

In the last years, research in Computational Linguistics has been relying heavily on textual corpora. From lexical acquisition to machine translation, most applications require large corpora that reflect a wide range of domains and genre. Certain tasks, on the other hand, like information retrieval, ontology and terminology extraction necessitate domain-specific corpora.

Many approaches in natural language processing make use of knowledge-poor resources, even raw corpora, while the rest need annotated textual databases which provide the necessary linguistic (grammatical, syntactic, or semantic) information. Manual annotation of such corpora is very costly and time-consuming, while using a set of linguistic tools in order to automatically provide annotation tags for the corpus tokens appears to be a challenge. Many languages, including Modern Greek, are not sufficiently equipped with linguistic resources, thus significantly reducing the options and the ability of language engineers for research and development.

In this paper we describe DELOS, a Modern Greek corpus of economic-domain consisting of 5 million word tokens. DELOS was created at first in order to be utilized in a lexicographic application for economic terminology and it has been further restructured to constitute a standardized, uniformly annotated textual database, easy to use for tasks reliant on text processing. Annotation includes morphological and shallow syntactic information such as phrase chunking and elementary subject-object relations.

Regarding morphology, Modern Greek is language rich in inflections. There are eleven different part-of-speech (pos) categories, six of which are declinable. Features characterizing nouns, articles and adjectives are three genders, four cases and two numbers. Verbs are characterized by their voice, mood, tense, number and person. Pronouns have different features depending on their type (personal, relative, interrogative etc.). As to sentence structure, Modern Greek is free in the order in which phrases are combined to form clauses. The subject and the object(s) of a verb can assume any position within the clause, either preceding or following the verb.

The rest of the paper is organized as follows: First the sources of the collection are presented and its coverage described. The linguistic tools used to automatically tag it are described in section 3. Statistical information concerning the corpus as well as a detailed explanation of the tagset are given in sections 4 and 5. The tools' accuracy and the evaluation of the linguistic process follow in section 6 and concluding remarks in the last section.

2. Makeup of DELOS

The motivation for creating DELOS was the DELOS project (Sintichakis et al., 2000), funded by the Greek Ministry of Development, which aimed at constructing a Greek-English dictionary of economic terminology. In order for the corpus to constitute a reliable source of information for lexicographic purposes, it had to possess balance and representativeness. Texts were sampled systematically to reflect tendencies in spoken and written language. The collection consists of texts taken from the financial newspaper EXPRESS, reports from the Foundation for Economic and Industrial Research, research papers from the Athens University of Economics and several reports from the Bank of Greece. The documents are of varying genre like press reportage, news, articles, interviews and scientific studies and cover all the basic areas of the economic domain, i.e. Microeconomics, Macroeconomics, International Economics, Finance, Business Administration, Economic History, Law and Economics, Public Economics etc.

3. Annotation Tools

As a first processing stage, numerous text filters were applied for the correction of typical typing errors and the conversion of character sets. The collection of texts was then linguistically annotated by a series of tools which operate on the GATE text engineering platform (Cunningham et al., 1996) and which are applied to the texts sequentially, each adding its own tagset to the corpus. Their output is in plain text as well as in XML format. A brief description of each of these modules follows next.

3.1. Tokenizer

The tokenizer detects the word tokens in the corpus, i.e. words, abbreviations, numbers, punctuation marks etc.

The tool is actually embodied within the sentence boundary detector.

3.2. Sentence Boundary Detection

Sentence boundaries were detected by a sentence splitter based on a variation of transformation-based error-driven learning (Stamatatos et al., 1999). Acronyms and abbreviations constitute a significant source of ambiguity regarding the detection of sentence boundaries. The splitter uses a small, manually annotated, training sub-corpus in order to extract a set of disambiguation rules, which are then applied to every potential sentence boundary in the actual corpus in order to determine sentence boundaries. Neither lexicons of any sort, nor lists of abbreviations are utilized.

3.3. Morphological Analysis

The next tool is a morphological analyzer (Sgarbas et al., 1998) developed according to the two-level morphology model. Rules specifying the permissible alterations to the orthographic form of a morpheme are realized as finite state transducers. A lexicon representing the morphotactic description of the language consists of a set of morphemes and information about the order in which these can be concatenated. Currently it consists of approximately 60,000 entries of lemmata. For the declinable words that are included in the morphological lexicon, the lemma, gender, number, case, person, mood values of the word are available as well as more specific information depending on its pos category and type information for pronouns, conjunctions etc. For words not covered by the lexicon, morphological information may be guessed based on the word suffix.

3.4. Phrase Chunking

The boundaries of intrasentential, non-overlapping noun, verb, prepositional, adverbial phrases and conjunctions are detected by a multi-pass chunker (Stamatatos et al., 2000). Simple chunks are identified during the first parsing passes while later ones deal with more complicated situations (i.e. combining already detected small chunks into longer ones). The chunker exploits minimal linguistic resources: a keyword lexicon containing 450 keywords (i.e. closed-class words such as articles, prepositions etc.) and a suffix lexicon of 300 of the most common word suffixes in Modern Greek.

3.4.1. Headword detection

Using empirical rules, the headword of the noun and prepositional phrases is detected as a last step of the chunker. The headword of a phrase is the word which determines the basic grammatical properties of the phrase (case, number etc.).

3.5. Shallow Parsing

The parser, last in the tool chain, detects elementary shallow syntactic dependencies, i.e. subject-verb-object relationships using the linguistic information provided by the previous modules along with a set of simple empirical rules taken from the Modern Greek syntactic theory. POS category information, case, voice and number information as well as information about the relative distance of the candidate subject or object from the verb are essential. A distinction between direct and indirect objects is made. Candidate subjects and objects are limited to being heads of noun and prepositional phrases as opposed to entire clauses.

4. Corpus Statistics

A detailed statistical description of the corpus is shown in the following table. Due to its economic domain, DELOS is rich in numbers (monetary amounts and percentages), proper nouns (names of people and companies), company acronyms and abbreviations. It consists of a total of approximately 5 million word tokens which form around 140,000 sentences.

Phrases	Noun	36,6%
	Verb	30,9%
	Prepositional	27%
	Adverbial	5,5%
Word tokens	Words	84,1%
	Punctuation marks	8,9%
	Abbreviations/Acronyms	3,3%
	Numbers	2,9%
	Other Symbols	0,8%
Words	In Greek alphabet	98,2%
	In Latin alphabet	1,8%
Words in Greek	Nouns	29,9%
	Verbs	10,2%
	Adjectives	10,6%
	Pronouns	3%
	Articles	15,2%
	Adverbs	5,8%
	Numerals	1,5%
	Conjunctions	6%
	Particles	1,6%
	Prepositions	9,2%
	Residuals	7%

Table 1: Statistical data for the corpus

By residuals we mean transliterated words (foreign words written in the Greek alphabet) and interjections.

5. Tagset Description

5.1. Morphological tagset

The entire morphological tagset appearing in DELOS is presented in this section. It was selected to be in accordance to the annotation scheme proposed in the PAROLE project (LE2 4017-10379) which was implemented manually for a small balanced corpus by the Institute for Language and Speech Processing (ILSP).

The tag format is

word<tag*lemma>

Word is the actual word as it appears in the corpus, *lemma* is its lemma, and *tag* is a set of characters providing morphological information for the word depending on the pos category of the word. The set of characters for every pos is shown in the following tables. Any invalid field or any field with unknown value has a dash (-) as a symbol.

Feature	Value	Symbol
Category	Verb	V
Type	Present participle	p
	else	-
Tense	Present	p
	Past	a
	Future	f
Person	1	1
	2	2
	3	3
Number	Singular	s
	Plural	p
Gender	Masculine	m
	Feminine	f
	Neutral	n
Voice	Active	a
	Passive	p
Case	Nominative	n
	Genitive	g
	Accusative	a

Table 2: Verbs

Examples:
εγκρίθηκαν<V--3p-p-*εγκρίνω>
διευκρίνισε<V--3s-a-*διευκρινίζω>

Feature	Value	Symbol
Category	Noun	N
Subcategory	Common	c
	Proper	p
Gender	Masculine	m
	Feminine	f
	Neutral	n
Number	Singular	s
	Plural	p
Case	Nominative	n
	Genitive	g
	Accusative	a

Table 3: Nouns

Examples:
μετοχών<N-fpg*μετοχή>
αποτελέσματα<N-npn*αποτέλεσμα>

Feature	Value	Symbol
Category	Adjective	A
Degree	Positive	p
	Comparative	c
	Superlative	s
Gender	Masculine	m
	Feminine	f
	Neutral	n
Number	Singular	s
	Plural	p
Case	Nominative	n
	Genitive	g
	Accusative	a

Table 4: Adjectives

Examples:
οικονομικά<A-npn*οικονομικός>
γαλακτοκομικών<A-npg*γαλακτοκομικός>

Feature	Value	Symbol
Category	Pronoun	P
Subcategory	Personal	p
	Relative	r
Person	1	1
	2	2
	3	3
Gender	Masculine	m
	Feminine	f
	Neutral	n
Number	Singular	s
	Plural	p
Case	Nominative	n
	Genitive	g
	Accusative	a

Table 5: Pronouns

Examples: αυτά<pp3npn*εγώ>

Feature	Value	Symbol
Category	Article	T
Subcategory	Definite	d
	Indefinite	i
Gender	Masculine	m
	Feminine	f
	Neutral	n
Number	Singular	s
	Plural	p
Case	Nominative	n
	Genitive	g
	Accusative	a

Table 6: Articles

Examples: των<T-mpg*ο>

Prepositional articles (*στον*, *στην* etc.) are tagged like articles except for the first character introducing their tag, which is S, for prepositions (see Table 11). Their lemma value is that of the article, i.e. *ο*. For example the tag of *στον* would be <*STdmsa*ο*>.

Feature	Value	Symbol
Category	Numeral	M
Subcategory	Cardinal	c
	Ordinal	o
Gender	Masculine	m
	Feminine	f
	Neutral	n
Number	Singular	s
	Plural	p
Case	Nominative	n
	Genitive	g
	Accusative	a

Table 7: Numerals

Examples: μία<Mcfsa-*ένας>

Feature	Value	Symbol
Category	Conjunction	C
Subcategory	Coordinating	c
	Subordinating	s

Table 8: Conjunctions

Examples: και<Cc*και>, αλλά<Cs*αλλά>

Feature	Value	Symbol
Category	Particle	U
Subcategory	Negation	n
	Future	f
	Subjunctive	u

Table 9: Particles

Examples: δεν<Un*δεν>, να<Uu*να>

Feature	Value	Symbol
Category	Residual	X
Subcategory	Foreign word	f
	Acronym	a
	Abbreviation	b

Table 10: Residuals

Examples: Nikas<Xf>, κ<Xb>, Π.Γ.<Xa>

Feature	Value	Symbol
Category	Adverb	R
	Preposition	S
	Interjection	I
	Punctuation mark	F

Table 11: Remaining POS tags

Examples: ειδικότερα<R*ειδικότερα>, για<S*για>, '<F>

5.2. Syntactic tagset

Noun, verb, adverbial, prepositional phrases and conjunctions linking them are introduced by the tags *NP*, *VP*, *ADP*, *PP* and *CON* respectively. The phrase body is enclosed within square brackets. The * symbol at the beginning of a word denotes a headword in a noun or a prepositional phrase. Below follows a piece of morphologically and syntactically annotated text.

*NP[*TH<T-fsa*ο> *διανομή<N-fsa*διανομή> μερίσματος<N-nsg*μέρισμα> αξίας<N-fsg*αξία> 90 <Mc---*90> δρχ<Xb>] PP[ανά<S*ανά> *μετοχή<N-fsa*μετοχή> ,<F>] VP[αποφάσισαν<V--3p-a-*αποφασίζω>] NP[οι<T-mpn*ο> *μέτοχοι<N-mpn*μέτοχος> της<T-fsg*ο> \<F> Π.Γ.<Xb> ΝΙΚΑΣ<Npfsn*unknown>] PP[κατά<S*κατά> τη<T-fsa*ο> χθεσινή<A-fsa*χθεσινός> πραγματοποιηθείσα<A-fsa*unknown> τακτική<A-fsa*τακτικός> γενική<A-fsa*γενικός> *συνέλευσή<N-fsa*unknown> τους<pp3mpg*εγώ> .<F>]*

The distribution of the dividend of a value of 90 drs per share was decided by the shareholders of P.C. NIKAS during the general conference held yesterday.

In the XML output format, every constituent (phrasal or lexical token) is given a unique id number. Verb phrases have a subject, direct and indirect object field that takes the value of the noun or prepositional phrase id of the phrase the headword of which functions as a subject, direct and indirect object of the verb phrase's verb respectively.

6. Evaluation of Tagging

Starting with the evaluation of the sentence boundary detection process, the detector reached an accuracy of 98.5%. As *accuracy* we define the percentage of the number of positive (a punctuation mark considered wrongly to be the end of a sentence) and negative (a sentence end not detected) errors. The detector has been trained by manually tagging the sentence boundaries of approximately 15000 sentences of the corpus and tested on 1500 new corpus sentences. Considering that, as mentioned before, DELOS is rich in proper nouns, abbreviations, acronyms, numbers and punctuation marks, and therefore a baseline accuracy value (i.e. regarding every full stop, exclamation mark, quotation mark and set of dots as a sentence boundary) does not exceed 57-58%, the accuracy reached is quite satisfactory.

The morphological lexicon at this time covers most of the closed-class words appearing in the corpus and approximately 65% of the declinable words and it is being constantly enriched. More specifically, 39% of the adjective lemmata, 25% of noun lemmata and 31% of the verb lemmata and most of the closed-class words appearing in the corpus are currently unknown. For declinable words included in the lexicon, accuracy of their morphological features exceeds 98%. For those words not covered in the lexicon, pos tagging performance turns out to be as shown in the following table.

POS	Precision (%)	Recall (%)
Verbs	83	92
Nouns	77	93
Adjectives	55	82
Adverbs	71	92
Pronouns	83	96

Table 12: Precision and recall for words not included in the lexicon

The recall and precision metrics for the morphological analyzer, the chunker and the shallow parser are defined as follows:

Recall = the number of correctly predicted types divided by the total number of types appearing in the input text.

Precision = the number of correctly predicted types divided by the total number of types predicted.

A type in the above definitions can be a pos category, a phrase boundary or a shallow parsing relation. The above results have been obtained by testing 1000 occurrences of every pos category (either predicted or theoretical, depending on whether positive or negative examples are searched for) in the corpus.

Nouns and adjectives have similar and sometimes identical endings. Adjectives and adverbs, as well as articles and certain types of pronouns have very often the same orthographic form. Their distinction is not straightforward without context information, which explains their lower precision values.

The tag correctness in respect to the most important grammatical features, again for words not included in the lexicon, given, however, that their pos tag is predicted correctly, is shown in Table 13.

Feature	Accuracy (%)
case	92
gender	86
number	94
person	82
voice	57
mood	57

Table 13: Accuracy of various features for words not included in the lexicon

For the chunker, testing reveals 89.5% recall and 94.5% precision, which is more than encouraging as Delos is a corpus rich in complicated syntactic structures. Foreign words, the endings of which are similar to endings of Modern Greek words, are also the cause of several chunking errors. Table 14 shows the performance of the chunker in more detail for every phrase type.

Phrase	Precision (%)	Recall (%)
NP	88,6	91,1
PP	99,3	93,3
VP	98,1	91,3
ADP	96,2	72,4

Table 14: Precision and recall for every phrase type

The problem of adverbs that have an identical orthographic form with adjectives accounts for the lower recall value in the detection of adverbial phrase boundaries. Lower precision in the noun phrase detection is attributed to the significant number of noun phrases which are not introduced by an article or a pronoun.

Recall and precision for the subject-verb-object dependencies detection are approximately 70%. These results have been obtained by testing 500 verb occurrences (a total of approximately 200 different verbs) in the corpus. Certain errors in previous stages (like case and number tagging of the headword of noun phrases) are to a large extent responsible for decrease in shallow parsing performance.

7. Conclusion

The problem of several languages not being sufficiently equipped with adequate resources is significant for language engineers. Full manual construction of such resources is expensive and not always feasible. In this paper we have presented the creation of an economic corpus for Modern Greek as well as its automatic annotation process and detailed description. Annotation reaches the level of shallow parsing. DELOS can be used for any linguistic application like extraction of economic terminology or ontological information.

8. References

Cunningham, H., Y. Wilks and R. Gaizauskas, 1996. GATE: A General Architecture for Text Engineering. In *Proceedings of the 16th International Conference on Computational Linguistics*, COLING '96, Copenhagen, 1057-1060.

Sgarbas, K., N. Fakotakis and G. Kokkinakis, 1998. A Morphological Description of Modern Greek using the Two-Level Model. (In Greek). Proc. Of the 19th Annual Workshop, Division of Linguistics, University of Thessaloniki, Greece, April 23-25, 419-433.

Sintichakis, M., K. Kermanidis and T. Z. Kalamboukis, 2000. Corpus Analysis for Applied Lexicography. Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries, COMLEX 2000, Kato Achaia, Greece, September 22-23, 121-126.

Stamatatos, S., N. Fakotakis and G. Kokkinakis, 1999. Automatic Extraction of Rules for Sentence Boundary Disambiguation. Proceedings of ACAI 99, Workshop on Machine Learning in Human Language Technology, 88-92.

Stamatatos, S., N. Fakotakis and G. Kokkinakis, 2000. A Practical Chunker for Unrestricted Text. Proceedings of the 2nd International Conference of Natural Language Processing, NLP 2000, 139-150.