" L e x i f a n i s "
A Lexical Analyzer of Modern Greek

Yannis Kotsanis - Yanis Maïstros
Computer Sc.Dpt. - National Tech. University
Heroon Polytechniou 9
GR - 157 73 - Athens, Greece

'l' écriture fait du savoir une fête' R.BARTHES

ABSTRACT

Lexifanis* is a Software Tool designed and implemented by the authors to analyze **Modern Greek Language** ('Δημοτική'). This system assigns **grammatical classes** (parts of speech) to 95-98% of the words of a text which is read and normalized by the computer.

By providing the system with the appropriate grammatical knowledge (i.e.: dictionaries of non-inflected words, affixation morphology and limited surface syntax rules) any "variant" of Modern Greek Language (dialect or idiom) can be processed.

In designing the system, special consideration is given to the Greek Language morphological characteristics, primarily to the inflection and the accentuation.

In Linguistics, Lexifanis, can assist the generation of indexes or lemmata; on the other hand readability or style analysis can be performed using this software as a basic component. In Word Processing this software may serve as a background to build dictionaries for a spelling checking and error detection package.

Through this study our research group has set the basis in designing an " expert system " which is intended to "understand" and process Modern Greek texts. Lexifanis is the first working tool for Modern Greek Language.

---

* 'Λεξιφάνης' : Who Brings the Words to Light. Name given by Lucian (circa 160 A.C.) to one of his dialogues.

PROLOGUE

In Linguistics the systematic identification of the word classes rises several questions in regard to the morphemic analysis. In Computational Linguistics several research areas use fundamental information such as the "word class" of a given word, isolated or in its context. In Computer Science the automatic processing of Greek texts is based on relevant knowledge, at the lexical level.

In an effort to present a software tool intended to **identify** the **grammatical classes** of the words we have designed and implemented Lexifanis. We have used modern greek texts as a testbed of our system, but Lexifanis, can process any "variant" of modern greek, and even ancient greek language, provided that it is appropriately initialized.

In this paper, whenever we use the term greek or greek language we refer to the **modern greek language** ('Δημοτική') in its recent **monotonic** version (i.e. a single accent is used, instead of three, and there are no breathings -'πνεύματα')

WORD CLASSES

We have found that **morphological analysis** of the greek words can provide adequate information for the word class assignment. The majority of the words in a text can be assigned a unique (single class). However, there exist some words that may be assigned two "possible" classes. This ambiguity is inherent to their morphology. On the other hand we know that consideration of the **words in their context** may disambiguate this classification, if required. In this work there is no need to use any stem dictionary.

The fundamental information used by Lexifanis to provide the classes of all greek words is extracted from the affixation morphology and especially from a **morphemic suffix analysis.** In this domain, we follow three axes of investigation : the "**Accentual Scheme**", the "**Ending**" and the "**Pre-ending**" of each word.

Accentual scheme

The "accentual scheme" of the word reflects the position of the stress on the word: The stress may come only on one of the last three syllables (law of the three syllables). This scheme is identified in our system by a code number. Table 1 lists all possible schemes and their corresponding identification codes (IC).

TABLE 1 : "accentual scheme" of the greek words

accent. scheme	IC	example	
! '	0	ð'	: will
!e	1	θα,πως	: will,that
!é	2	πώς(;)	: what(?)
!eé	3	παιδί	: child
!ée	4	χάρη	: grace
eeé	5	αρχαϊκός	: archaic
eée	6	συνθέτω	: I compose
ée	7	πρόβλημα	: problem

Notation

!	"word start delimiter"
e	"syllable"
'	"accent"
,	"apostroph"

An example to illustrate the above feature is the following:

δι-και-ο-σύ-νη	(:justice)	IC=6	NOUN
χαρ-μό-συ-νη	(:joyful)	IC=7	ADJ

Ending

A detailed suffix analysis of the highly inflected greek language [KOYP,67] [MIRA,59] indicates that there exist morphemes at the end of the word which can be used to identify the grammatical classes of the words.

The morphological analysis, presented in this paper,is based on a right-to-left scanning of the words. This analysis identifies word **suffixes**, named hence-fourth **endings.** These endings may not necessarily coincide with the **inflectional suffixes,** described in the greek grammar [TRIA,41]. Consider for example the following pair of words highlighting the difference in the ending of the two words. (In this example the ending is the inflexional suffix, as well).

εκτέλ - εσ - η (: execution)	NOUN
εκτέλ - εσ - α (: I have executed)	ADJ

Notice the identical accentual scheme of the above two words.

Pre-ending

On the other hand, these endings reflect the incidental cases of morphemic ambiguity [KOKT,85] in the inflectional greek language. This ambiguity can be resolved if we further penetrate to the word to identify what we call **pre-ending.** This pre-ending, in most cases, can be easily used to disambiguate word classes and it yields to a unique class assignment when the ending alone is not sufficient. Generally, the pre-ending does not coincide with the **derivational suffix** of the word under consideration [TPIA,41].

Let us now consider the following example :

κάν - ατε	(: you have done)
θάνατ - ε	(: death, in vocative case)

where,the consideration of the linguistic inflectional sufixes -ατε and-ε are completely misleading, as far as the class assignment is concerned. You may notice that these two words have the same pre-ending -ατ-. In this case a further morphemic penetration in the word is required to resolve the ambiguity [KRAU, 81]:

κάν - ατ - ε	VERB
θάν - ατ - ε	NOUN

The morphemes identified at this last penetration may not necessarily form the stem of these words. Our system classifies the first word as a verb and the second as a noun.

Words in their Context

Finally, if more ambiguities exist in word class assignment, a consideration of the "words in their context" may be added to the affixation morphology. This classification technique is fruitful in poorely inflectional languages, such as English [CHER,80], [KRAU,81], [ROBI,82].

This syntax analysis is recommended when the task is to determine the classes of the words in a **whole text**, as opposed to the class assignment to **isolated words.** By this analysis we gain information from up to two words that precede or follow the word under classification [TZAP,53]. The following is a classic disambiguation example :

οι αντιθέσ - εις (: the contrasts) NOUN
να αντιθέσ - εις (: to contrast) VERB

IMPLEMENTATION

Dictionaries of Non-Inflected Words

Greek language is highly inflected. However, due to the fact that one out of two words of a text is a non-inflected word we have constructed the **dictionaries of non-inflected words** containing about 400 entries. In these dictionaries we accommodated all the non inflected words, that have no derivational suffix, of modern greek, such as **particles**, pronouns, prepositions, conjunctions, homonyms,etc. and the inflected articles.

Each word that enters Lexifanis is first searched in these dictionaries. If there exist an identical entry, its class is assigned to this word. Fig. 1 lists some of the entries of these dictionaries. As an example consider "στο" (:to the,it). This word can be either "article with preposion" or "pronoun".

```
        art :  η ο οι των
   art_pron :  τη της του ...
   art.prep :  στης στου στων
art.prep_pron : στη στο στα ...
   prep_pron : με μ' σε σ'
       pron :  μου μας εμένα ...
       prep :  από χια προς ...
       conj :  ή και αλλά ...
    homonym :  που μην να ...
   particle :  ας θα θ'
        num :  δύο δυο τρεις ...
        adv :  πού χεια χθες ...
```

Fig. 1 Part of the Dictionaries of Non-Inflected Words

Morphological Analysis

The Morphological Analysis is performed using about 250 rules. The user may add, delete or modify anyone of these rules. These rules contain all the information relevant to the endings and pre-endings. During this phase, the inflected words, mainly verbs and nouns, are identified. Efficient search is carried out using the accentual code, mentioned above.

EXAMPLE: "Five" Morphological Rules :

```
                <!έΞ/eΞ> <n/ης>   : noun
                   <eΞ> <α/ες/ε> : verb
<μέταΞ/δοΞ/άτοΞ/
    όδεΞ/όλεΞ/όρεΞ>        <α/ες/ε> : name
<άμαΞ>                    <α/ας/ες>: noun
<αμαΞ>                    <ών>     : noun
```

Notation

!	"word start delimiter"
e	"syllable"
'	"accent"
/	"exlusive or"

Limited Syntax Analysis

When we want to analyze and classify the words of a text as a whole, Lexifanis examines the word under consideration in its context. This can be accomplished by invoking the nearly 25 **Limited Surface Syntax Rules.**

This step is recommended, in case a word, is assigned two possible classes (double class assignment), see Table 2, using only the affixation morphology. This double class assignment is due to the ambiguity inherent to the morphology of the word.

EXAMPLE: "Two" of the limited surface syntax rules :

```
<prep_pron> <verb>
                  => <pron> <verb>

<prep_pron> <art_pron>  <unclass>
                  => <prep> <art> <name>
```

THE SOFTWARE SYSTEM

Lexifanis is a set of structured programms implemented in two versions :

* The BATCH system, assigns classes to the words of a whole text. This system performs the limited syntax, mentioned above, in addition to the morphology.

* The INTERACTIVE system, assigns classes to isolated words. This system performs only the morphological analysis.

Structure of Lexifanis

The whole software system is designed and implemented in MODULES or PHASES, the structure of which is illustrated in the

Block Diagram of the Figure 2. The description of each module follows.

INITIALIZATION - During this phase two processes take place :

* the creation of the Dictionaries of Non-Inflected Words, and
* the generation of the appropriate Automata required to express the morphological rules and the surface syntax rules

INPUT AND NORMALIZATION OF THE TEXT- The interactive version of the software system performs only the accentual scheme process, whereas the batch version performs this process in parallel to the input and normalization processes. Normalization or Word Recognition is the task of identifying what constitutes a word in a stream of characters.

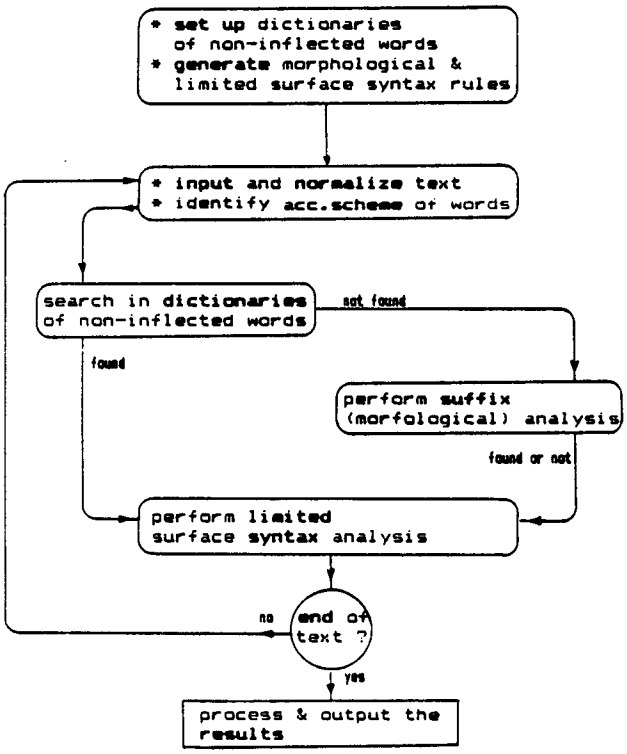

Fig. 2 Structure of Lexifanis

SEARCH IN DICTIONARIES - All the Non-Inflected Words, with the same accentual scheme, and word length, are grouped together forming a set of small dictionary-trees, "cultivated in a two dimentional...garden", minimizing thus the search time (Fig.3).

SUFFIX ANALYSIS - This is the main process of our system which is activated for words not contained in dictionaries. Finite State Automata [AHO ,79] are used to represent the morphological rules.

LIMITED SYNTAX ANALYSIS - The relevant information is represented by automata.

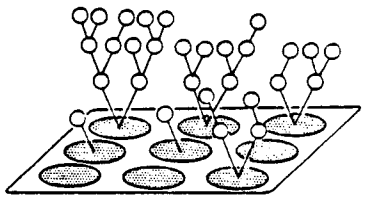

Fig. 3 the ... two dimentional garden

RESULTS - This module is best fitted to the batch version of our system, but it can be used in the interactive version, as well.

TABLE 2 : Results obtained from a Scientific Text

	after morph. analys. %	after surface syntax %
single classes		
1. article	5.16	13.53
2. article with prepos.	0.00	1.20
3. pronoun	5.11	6.42
4. numeral	3.91	3.91
5. preposition	2.96	5.26
6. conjuction	6.47	8.22
7. adverb	6.12	6.12
8. particle	0.60	0.70
9. noun	12.73	12.98
10. proper noun	0.30	0.30
11. adjective	7.27	7.27
12. participle	1.50	1.50
13. verb	13.18	13.18
	65.31	**80.60**
double classes		
14. art_pronoun	11.78	2.16
15. art with prep_pron	1.25	0.00
16. preposition_pronoun	2.36	0.05
17. non-inflected homonym	2.71	0.85
18. name : noun_adject	11.33	11.33
19. adject_adverb	2.06	1.80
	31.48	**16.69**
unclassified words	**3.21**	**2.71**

The Results concerning the classification of a greek text, are summarized in Table 2.

* A single class is assigned to 80-90% of the words of any text, 8-15% are assigned two possible classes (double class assignment),and the remaining 2-5% of the words, are left unclassified.

* The variation of the above percentages is due to the difference in style of the texts being processed. A scientific writing, for example, contain fewer ambiguities than a poem.

COMPUTATIONAL DETAILS

Lexifanis' modules are written in "Pascal" programming language. This software runs under NOS operating system on a Cyber 171 main frame computer. Top-down design and structured programming guarantee the portability of this product.

The system uses about 35 Kilowords of the Cyber computer memory (60bits/word) and it requires 12 seconds "compilation time". The batch version classifies the words at a rate of 110 word classes per second.

APPLICATIONS

Lexifanis is a complete software tool which assigns classes to isolated words entered by the user or, alternatively, to all the words of an input text. This system can be useful to a variety of applications, some of which are listed below. The modularity in its design and implementation, along with the generality of the concepts implemented guarantee a property to our system : it can be easily integrated into various software systems.

The most apparent application of Lexifanis is, in Lexicography, the generation of "morpheme-based" dictionaries and the generation of lemmata.

Lexifanis may serve as a background in a spelling checking and error detection package , or any "writers aid" software system.

Finally, Machine Translation would be another major area of application where Lexifanis may be included, as a module or process, in an "expert system".

EPILOGUE

... we have presented a software tool, which assigns grammatical classes to the 95-98% of the words of a given text.

This system performs suffix analysis to assign classes to all the greek words. For the first time accentual scheme has been proved useful in the classification of greek words. Moreover, ambiguities inherent to the suffix morphology of greek words can be resolved without any stem dictionary ...

REFERENCES

[ΚΟΥΡ,67] : Γ. Κουρμούλη, Αντίστροφον Λεξικόν της Νέας Ελληνικής,Αθήνα,1967

[ΤΖΑΡ,53] : Α. Τζάρτζανος, Νεοελληνική Σύνταξις, 2 τόμοι, Αθήνα, 1946/1953

[ΤΡΙΑ,41] : Μ. Α. Τριανταφυλλίδης, Νεοελληνική Γραμματική, Αθήνα 1941/1978

[AHO ,79] : A.Aho, Pattern Matching in Strings, Symposium on Formal Language Theory, Santa Barbara, Univ. of California, Dec. 1979

[CHER,80] : L.L.Cherry, PARTS-A System for Assigning Word Classes to English Text, Computing Science Technical Report #81, Bell Laboratories, Murray Hill NJ 07974, 1980

[KOKT,85] : Eva Koctova, Towards a New Type of Morphemic Analysis, ACL, 2nd European Chapter, Geneva, 1985

[KRAU,81] : W.Krause and G.Willée, Lemmatizing German Newspaper Texts with the Aid of an Algorithm, Computers and the Humanities 15, 1981

[MIRA,59] : A . Mirambel, La Langue Grecque Moderne - Description et Analyse, Klincksieck, Paris, 1959

[ROBI,82] : J.J.Robinson, DIAGRAM : A Grammar for Dialogues, Comm. of the ACM, Vol.25, No 1, 1982

[SOME,80] : H.L.Somers, Brief Description and User Manual, Institut pour les Etudes Sémantiques et Cognitives, Working Paper #41, 1980

[TURB,81] : T. N. Turba, Checking for Spelling and Typographical Errors in Computer-Based Text, Proceedings of the ACM SIGPLAN-SIGOA on Text Manipulation, Portland - Oregon, 1981

[WINO,83] : T. Winograd, Language as a Cognitive Process, Vol. I : Syntax, Addison - Wesley, 1983