A COMPUTATIONAL MODEL FOR THE MORPHOLOGICAL ANALYSIS OF THE GREEK NOUN CATEGORY

ABSTRACT

This paper presents the first stage of research for the development of an automatic lemmatisation system for Modern Greek. The noun category of the language has been examined so far. The objective of the study was to identify - in the form of a suffix database - all the suffixes which provide single grammatical information about the lexical item they mark. The novelty of the approach lies first in what has been treated as suffix and, secondly, in the attempt to grammatically disambiguate words in a text on the basis of morphology alone, reducing dictionary matching to its minimum. The term **graphological groupings** (G.G.) proved the most appropriate for the suffixes of our database, as they range from one to seven characters and they cover in many cases both the inflectional and derivational suffixes of the language. Six hundred and fifty one graphological groupings (G.G.) out of a total of eight hundred and fifty have been identified as specifying, rendering dictionary matching redundant. The remaining one hundred and ninety nine provide two to three grammatical alternatives in the form of Gender/Number/Case (G.N.C.) combinations, requiring either dictionary matching or interactivity facilities for the disambiguation process. This database, supplemented by a number of simple transformation rules, gave birth to a program which achieves automatic lemmatisation for the majority of Greek nouns. The lemmatiser on the whole can be used :

a. for teaching Modern Greek grammar to the first classes of Primary School.

b. as a filter:

i) during Modern Greek computerised dictionary consultation for foreign learners of the language

ii) on the morphological analysis level for Machine Translation systems

iii) in automatic spelling-checking systems.

INTRODUCTION

Most of the work done in Computational Linguistics and natural language processing over the last 30 years has concentrated mainly on the analysis of Western European languages and of Japanese, Chinese or Arabic. It is relatively recently - during the last decade - that a special interest has been shown by Greek linguists and computer scientists in examining their own language with computational methods and in attempting to apply such strategies and ideas on Modern Greek.

The initial idea of the research was to implement a lemmatiser that can offer grammatical disambiguation of lexical items by means of their inflectional markers alone; that is to say, without any dictionary consultation. This approach has proved to work with good results for Spanish and French in particular. As regards Greek, although the multiformity of the morphology and grammar encountered in Ancient Greek has become subject - sometimes prescriptively - to considerable simplification, the language still employs an extremely large number of morphological features to realise grammatical notions.

Often it is a one-to-one relation, a unique morpheme, such as {που}, {εδώ}, {τώρα}, for a particular notion. Nevertheless, in most cases the situation is more complex. On the one hand, there are many instances in which a single morphological feature marks more than one grammatical category. For example the inflectional suffix *-είς* is common for nouns, i.e. *τομ-είς* (masc.,plur.); adjectives, i.e. *πολυπληθ-είς* (masc./fem.,plur.) and verbs, i.e. *περιποιηθ-είς* (2nd pers.,sing.,indic/subj.). The suffix *-ά,* is common for adjectives, i.e. *πλατι-ά* (neut.,plur.); nouns, i.e. *φορ-ά* (fem.,sing.); adverbs, i.e. *σωστ-ά* and verbs, i.e. *αγαπ-ά (*3rd pers.,sing.,indic.).

On the other hand, due to the constant evolution of the language from "katharevousa" to "demotic" Greek, there are many cases in which a number of free allomorphs conveys a specific concept, as is the case with the morpheme of 3rd person, plural, past, progressive, passive. It consists of 3 allomorphs: - *ονταν*

- *όντανε*
- *όντουσαν* (1)

Moreover, the accent system of the language influences the assignment of lexical and grammatical meaning to a word as well , as is the case for instance with the word *πορτοκαλια.* If the accent is on the last syllable: *πορτοκαλιά,* then it refers to an orange tree or the orange colour of, say, a blouse. But with the accent on the penultimate syllable: *πορτοκάλια,* it refers to the fruits of such tree.

Lastly, within the same grammatical category, it is often very difficult to define the grammatical meaning of a member of this category based only on its inflectional marker. The suffix *-είς,* for example, can mark either subjunctive or indicative mood, and *-ους* can be the inflectional marker for:

(i) masc., plur., accus. : i.e. *λα-ούς (peoples)*
(ii) masc., sing., nom. : i.e. *παππ-ούς (grandfather)*
(iii) fem., plur., accus. : i.e. *οδ-ούς (street)*
(iv) fem., sing., gen. : i.e. *αλεπ-ούς (fox)*
(v) neut., sing., gen. : i.e. *δάσ-ους (forest)*
(vi) masc-fem., plur., accus. : i.e. *συζύγ-ους (husband-wife)*

Nevertheless, the main morphological marker system for the Greek language is the inflectional. Such a complex system that, at first glance, fails to define single grammatical properties, grammatical categories, or even members of a grammatical category as such, is an area urging systematic research.

In this light, the research initially aspired to determine all the inflectional markers that provide single grammatical information about the lexical item they mark. However, if all inflecting grammatical categories had been examined, the clarity and even the feasibility of the research might have been jeopardised. Six inflecting grammatical categories of Modern Greek, or even five along Quirk and Greenbaum's OPEN-CLASS lines (2), would undoubtedly have been an immense corpus for an initial examination.

Therefore, only one grammatical category was examined in isolation, the NOUN category (3), which was expected to function as the base for further expansion at a later stage. The aim of the research was thus limited to:

a) determining all the inflectional markers within the noun category

b) isolating those that function as unique morphemes and are consequently grammatically unambiguous

c) grouping the remaining ambiguous markers along with the alternatives of grammatical information each one conveys.

EARLIER APPROACHES

The usual approach implemented so far in natural language analysis of European languages has been to furnish each entry in the dictionary with the corresponding grammatical information codes (in terms of grammatical category, conjugations, declensions etc.) and the paradigmatic (inflectional) tables to which this information refers.

This is the case for instance with the technique adopted by Machine Translation systems for the grammatical disambiguation of sentences on the morphological level. SYSTRAN, for example, uses a number of morphological programs for the analysis of French which aim at determining whether a given ending is valid with a given type of noun/adjective/verb. Once an ending is valid for a specific grammatical category member, then this lexical item is automatically assigned its grammatical values.

For Modern Greek in particular, on the analysis level, EUROTRA followed a similar approach: dictionary of stems, inflectional tables and valid combinations of the two for the grammatical recognition of a word in a text.

For French, however, the Scientific Centre of IBM France attempted a different approach, similar to the one proposed by this study but of a narrower scope. By examining the 5 last characters of 150,000 French words taken from texts of the XIX and XX centuries, they succeded in establishing 1,172 "morphophonological" rules on the basis of which the exact "classe syntaxique" , as they call it, of 131,078 words can be recognised. They failed to distinguish between adjectives and nouns, but they were successful in distinguishing between adjectives/nouns on the one hand and verbs on the other. Grammatical tagging was thus

achieved, but not automatic lemmatisation or recognition of the subtler grammatical meaning of the words.

DESCRIPTION OF THE PROPOSED MODEL

The approach proposed hereby does not wish to break new ground with regard to the traditional computational approach to analysis problems of inflectional languages. The stages of initial scanning for a limited number of unambiguous lexical items or the syntactic analysis of a sentence after the morphological disambiguation of its items cannot disappear from our approach.

The novelty of the lemmatiser lies, however, on what was treated as a morph. As it has already been stated, there are inflectional suffixes in Modern Greek that occur in all inflectional tables of the noun category, regardless of gender, such as *-α* or *-ων*. Some others, i.e. *-ας* or *-ες* are particularly common for masculine and feminine nouns and others, i.e. *-ου* are particularly common for masculine and neuter.

However, it was observed that the longer the ending is, the smaller the number of nouns that end in this way is. Therefore, if we examined longer suffixes than the traditional ones, for example *-ούληδες, -οτών,* or *-ιδα,* it would be feasible to trace all the nouns that form the specific ending, classify them according to gender and reduce the number of ambiguous cases by classifying the smaller gender groups as exceptions, thus rendering dictionary consultation processes for the grammatical recognition of lexical items redundant.

This is what was actually attempted. It should be noted, however, that a different term was adopted for the endings examined, due to the fact that those covered by the research do not coincide with morphemes, morphs or even endings in the technical sense of the words. On the contrary, in most cases they are longer than the traditional inflectional suffixes, in many cases easily identified as derivational or even longer. The term used is GRAPHOLOGICAL GROUPINGS (G.G.) and it has proved to serve our purposes very well.

The examination of all possible G.G. gave positive results. There is indeed a large number of G.G. (651) that provide a unique piece of grammatical information about a noun in the form of a Gender/Number/Case (G.N.C.) pattern. The rest of the G.G. examined (199) provide more than one G.N.C. pattern: one hundred and seventy five G.G. offer two alternatives and twenty four G.G. offer three alternatives. These G.G. were not excluded from the study, as each of the G.N.C. patterns is fixed and can be treated as if it were the sole one. These encouraging findings led to one further step, that is the formulation of the necessary rules that can produce the Basic G.G. form of any G.G. included.

Thus, the product of the study consists of:

(1) a list of all possible G.G. of Modern Greek nouns, each one accompanied by examples and classified in 2 extensive sections, the SPECIFYING section and the NON-SPECIFYING section, presented in reverse alphabetical order

(2) the grammatical information conveyed by each G.G. in the form of a G.N.C. pattern

(3) a list of exceptions to each G.G. (if any)

(4) a set of rules which can help move from any number and case other than singular nominative back to the Base Form of the noun (that is, Singular Nominative).

A typical entry, in other words, is of the form:

```
                    ---------  G.N.C.
G.G.----------------------  (exceptions)
                    ---------  (rules)
```

For example :

```
                    -------  F.S.G.
                             i.e. κερασιάς, φορεσιάς, εργασίας, θυσίας etc.
-σιας  ----------  exceptions:  σωσσίας, αντιρρησίας, Μεσσίας
                                          (M.-F.,S.,N.)        (M.,S.,N.)
                    -------  *DELETE -ς FROM  -σιας
```

There are three basic commands, CHANGE, ADD and DELETE, with an additional MOVE command which allows for the shift of the accent to another syllable, whenever necessary. Each command can be used separately or in combination with any of the others, resulting in 1011 occurences.

The exception words were also given their G.N.C. pattern, apparently different from the G.N.C. pattern under examination and presented in brackets, so as to indicate on what grounds the specific nouns were classified as exceptions. However, it was not attempted to supply their Base Form rules as well, in order to avoid both overloading the "four boxes" layout and endangering the clarity of the information.

An additional list of ONLY PLURAL nouns was inserted before the exception word list for certain G.PLURAL.C patterns which presents nouns of the particular G.N.C. pattern that do not appear in their singular forms. There are not, for instance, Modern Greek words such as **λόγι* or **βράχι* deriving from *λόγια (words)* or *βράχια (rocks)*. Therefore, a command had to be implemented that would block the Base Form rules from applying. At that stage the only option was to present explicitly all nouns that fall into this category. In a future development of the project, however, a more economic technique must be devised.

Finally, a concise list of all the G.G. was included at the end of the analytical presentation of the corpus, where the G.G. were divided into 8 groups according to their final letter (A,E,H,I,N,O,Σ,Y). Each G.G. was then given the grammatical meaning(s) it conveys together with the number of nouns counted in the Reverse Dictionary of Modern Greek (Kourmoulis, 1967) that fall within each pattern. For example, the G.G. *-ευτών* marks 109 masculine nouns of genitive plural and 1 feminine noun of the same number and case : *-ευτών :* M.P.G. (109) / F.P.G.(1) In this way, frequency figures for the various forms of Modern Greek nouns were provided, without any attempt, however, of supplying frequency rates as well, as they will be automatically modified as soon as further grammatical categories are included in the study.

LINGUISTIC AND METHODOLOGICAL ASSUMPTIONS

LINGUISTIC

a) **Accent patterns**

It became obvious from the beginning of the research that the position of the accent could function as the sole distinctive morphological marker between two otherwise identical G.G. As has already been demonstrated with the example of *πορτοκάλια* vs *πορτοκαλιά*, there are numerous pairs of lexical items which convey different lexical and grammatical meanings according to the position of their accent. In the light of these "tonal paronyms", as M. Triantafillidis characterises them for reasons of similarity both in form and pronunciation *(θόλος-θολός, κάμαρα-καμάρα)* (4), the consideration of the various accent patterns of the examined G.G. seemed very promising.

Indeed, had the accent been ignored, we would have ended up with a large number of G.G. which mark various grammatical categories, each category consisting of large subgroups of various grammatical properties. Let us consider one very common Modern Greek G.G.: *-κια*. If the position of the accent is not taken into consideration, the word examined might be one of the:

(i) 158 feminine nouns, sing., nom.-accus.-voc.
(ii) 10 masculine nouns, sing., gen.-accus.-voc.
(iii) 1150 neuter nouns, plural, nom.-accus.,-voc.
(iv) 3040 feminine adjectives, sing., nom.-accus.-voc.
(v) 18 adjectives, feminine, sing., nom.-accus.-voc.
or
neuter, plural, nom.-accus.-voc.
(vi) the adverb *αντρίκια*

If, however, we examine the G.G. three times, according to the three possible accent patterns, we can firstly narrow down the number of possible grammatical meanings conveyed by each version of the G.G. and secondly reduce the actual number of lexical items that fall into each subcategory. Therefore, we have:

	------	64 fem., S.N.A.V.	i.e. *συκιά*
NOUNS -----	------	2 masc., S.G.A.V.	i.e. *Πλακιά*
	------	10 neut., P.N.A.V.	i.e. *σακκιά*

-κιά ----------

ADJECTIVES ---------- 3040 fem., S.N.A.V. i.e. *μαλακιά*

NOUNS

-κία ---------- 92 fem., S.N.A.V. i.e. *κακία*
(with the exception of 3 neuter
P.N.A.V. *σαρκία, δισκία, κοκκία*)

------ 1137 neut., P.N.A.V. i.e. *σκυλάκια*
-----NOUNS ---------- 2 fem., S.N.A.V. i.e. *φώκια*
------ 8 masc., S.G.A.V. i.e. *γυναικάκια*

΄-κια -- ADJECTIVES ---------- 18 fem., S.N.A.V./ neut., P.N.A.V.
i.e. *επινίκια*

--- ADVERBS ---------- 1 *αντρίκια*

The figures which precede each subcategory should not be taken as absolute. They merely represent the number of words G. Kourmoulis and his assistants selected in 1967 for the Reverse Dictionary of Modern Greek, the corpus of which served as the basic statistical tool for our research. They should, however, be considered as indicative of the occurence frequency of each G.G.

Unfortunately, the elimination of the circumflex symbol from the accent system of the language complicated our approach to the extent that nouns which - on the basis of this symbol - could immediately be recognised as such, or even be distinguished in terms of G.N.C. pattern, lost their one and only distinctive feature. Nouns, for example, ending in *-εια* would be definitely recognised as neuter, plural, N.A.V. when marked with the circumflex symbol, and as feminine, singular, N.A.V. when marked with the accute (cmp. *πορ-εία* and *καφεν-εια*). Unfortunately one accent symbol had to be used throughout the study and such G.Gs. had to be classified in the NON-SPECIFYING section.

Bearing the Greek accent system rules in mind, the examination showed that:

(i) some of the G.G. are definitely specific of a certain G.N.C. pattern, regardless of the presence of an accent marker, i.e. *-μοι (πόλε-μοι, γά-μοι)*
(ii) others occur with a fixed accent position, i.e. *-ίτιδα (φλεβ-ίτιδα)*
(iii) there are cases in which the accent can either be part of the G.G. or immediately precede it, i.e. *΄-άμα (κλ-άμα, αμάλγ-αμα)*
(iv) there are cases in which the G.G. is never marked, the accent being on either of the two syllables preceding it, i.e. *΄-ση (δύ-ση, βάφτι-ση)*

To sum up, there are four possible accent patterns : -αβ
-άβ -αβ΄
΄-αβ -άβ
΄-αβ

(b) **Rare Number/Case patterns**

One of the data collection stages was the formulation of all possible G.G. which would derive from the Basic G.G. form. At that stage we came across certain nouns that do not form a particular Number/Case pattern in their everyday use. Should these nouns, and perhaps the G.G. to which they correspond, be handled descriptively and excluded from the study? Or should these nouns be included in the study in the form they would have prescriptively?

The creative aspect of the language provided us once again with the answer. The fact that certain forms of words are not used in the Modern Greek vocabulary does not mean that they cannot be created and used, if necessary. Therefore, the G.Gs. in question were retained and exemplified by nouns which, if formed, would end in a G.G. identical to the one classified.

(c) **Accent pattern of rare Number/Case patterns**

A relevant problem was the decision to be taken on the accent pattern of such "rare" words and Number/Case patterns. As already mentioned, there is still confusion sparked off by the co-existence in the Modern Greek vocabulary of words originating in Ancient Greek and katharevousa with words that are a clear product of contemporary use or loanwords.

As regards the first type of words (*ελπίδα, θάλασσα, κανόνας, γραμματέας* etc.), there are two main factors which determine their form and consequently the form in which they appeared in the study: a) a large number of grammatical rules which govern their use and form and b) the establishment of these forms in the native speakers' linguistic consciousness through extensive use. The second group (*κουκλάρα, παλιατζίδικο, μαγιό, κοιλαράς* etc.), however, often refuses to follow the morphological rules which apply in the first group, especially in spoken language.

Under these circumstances, the problem was how to handle nouns which belong to the two different groups but end in the same G.G., the main issue being the different position of the accent.

For the first group there are specific grammatical rules which dictate the shift of the accent when forming certain non-basic forms. Feminine nouns in plural genitive case, for example - apart from the ones that originate in the 3rd declension of the Ancient Greek nouns - shift the accent to the last syllable (*γυναικών, θαλασσών* etc.). Certain **Modern** Greek feminine nouns, however, (*κολυμβήθρα, φωνούλα, γυναικάρα* etc.) form plural genitive cases very rarely and, if at all, only in the spoken language. When they are formed, there is a great tendency among native speakers to retain the accent on the same syllable, the penultimate, breaking the aforementioned rule. The phenomenon on the whole has been explained as a gradual replacement of the plural genitive case in the everyday use of the language by alternative syntactic patterns such as prepositions + accusative case.

After long contemplation and interviews with native speakers, it was decided to act prescriptively and include G.G. and nouns in the form which specific rules dictate. The main argument for this approach was that the product of this study would most probably be used for written texts and checking procedures and should therefore function as a point of reference for grammatical accuracy as well. Only well established forms such as *κοκκινίλων* or *μαρμελάδων* were allowed in the corpus in their rule-breaking form.

d) **Allomorphs of G.G. forms**

We also came across a large number of nouns with two possible versions of the same G.G., the one easily identified as a remnant of the Ancient Greek and katharevousa spelling norm, and the other as the "modern" version of it i.e. *-ις /-η, -ότης /-ότητα*.

The initial idea was to exclude the older suffixes. Certain forms, however, still occur in the everyday language, books and press, some being used often especially in cliche expressions. In anticipation of such expressions in texts, as well as for reasons of completeness, it was decided to include both in the corpus.

Nevertheless, certain older G.G. forms such as *-ων*, occuring at the end of nouns encountered in the Reverse Dictionary of 1967 but not in the Modern Greek vocabulary, had to be modified and adjusted to their contemporary form, i.e. *-ωνας* for masculine entries and *΄-ώνα* or *΄-όνα* for feminine.

METHODOLOGICAL

a) **Classification methods**

As mentioned earlier the findings of the research were divided into two extensive sections. Section 1 consists of the specifying G.Gs. which convey only one kind of grammatical information and Section 2 consists of the non-specifying G.Gs. which provide alternative kinds of grammatical information, always in the form of a Gender/Number/Case pattern.

Two alternative classification methods were considered:

1) Subdivide the specifying section into three subclasses according to gender, retaining the reverse alphabetical order.

2) Subdivide the non-specifying section into subclasses according to the number of alternatives each G.G. provides and formulate subclasses of the same gender options; for example, all G.G. marking feminine and neuter nouns could be placed together.

Nevertheless, creating so many subclasses would make the result extremely complicated, far less economic than it is already. On the contrary, more compact presentation formats can already be envisaged, even an attempt to create what is known as a finite state automaton. Both subclassification ideas were soon abandoned. Given the fact that there is never a choice of more than two places to look for a G.G. (either in the specifying or the non-specifying section), it is expected that the reverse alphabetical order in which they were listed will suffice for a quick and effective consultation of the database.

b) **Exception lists**

For most of the G.G., whether specifying or not, the examination gave a number of nouns which do not agree with the majority of the nouns falling under a G.G. in terms of G.N.C. patterns. There are, for example, six nouns ending in *-ήνα: λειχήνα, χήνα, σφήνα, σπλήνα, σωλήνα,* and *σειρήνα,* which are of feminine gender, singular number and N.A.V. case, and not of masculine gender and G.A.V. case as are the majority of similarly ending nouns. Such sets were listed as exceptions.

The exception lists often include a variety of exceptional nouns in terms of G.N.C. patterns. For a specifying G.G. there were up to four different G.N.C. patterns identified, presented explicitly in the form

a. .....(M., S., G.)
b. .....(Nt., P., N.A.V.) etc.

Where an exceptional noun was found to constitute the 2nd part of a compound noun, for reasons of economy this noun was immediately repeated after the specific exception group, numbered as 2x and preceded by "-". An example of this case is 2a:-*παρέα* (F., S., N.A.V.) as an exception to the non-specifying G.G. *-έα* which marks nouns of masculine or masculine/feminine gender.

Interestingly enough, in some cases it was not complete nouns which functioned in such a way, but what should rather be called subG.G. A subG.G. is a graphological grouping which is longer than the basic one and occurs in a limited graphological environment. Consider, for example, the G.G. *΄-φωνα* (-phone) and *΄-γωνα*, subG.G. to the basic G.G. *΄-ώνα*. The problem which arose with the specific G.Gs. was that although they were morphologically limited and therefore easily recognised and assessed, they could not be classified as exceptions 2x, because they presented a high occurence percentage: 33,33%. Were meaning employed in the analysis, the problem could easily have been resolved. Yet, computers must depend on non-semantic criteria to identify all the *-phone* compounds of a language. The basic G.G. *΄-ώνα* had thus to be transferred to the non-specifying section, offering two alternatives: 69 masculine, singular, G.A.V. nouns and 40 neuter, plural, N.A.V. (the *΄-φωνα,΄-γωνα* nouns), leaving the 11 feminine, singular, N.A.V. nouns as exceptions.

It is hoped that all exceptions for all G.G. were included. Nevertheless, further research might reveal that there are more to be added. If we wished to cover idiomatic, dialectic, field specific or even slang expressions, we would be forced to increase exception lists as well. Lexicographers and future users of the database are the ones to ultimately decide on such expansions.

COMPUTATIONAL IMPLEMENTATION OF THE PROPOSED APPROACH

The suggested computational implementation of the produced database in 1989 was the following:

(1) **G.G. recognition**

Given a certain noun as input, whether in isolation or in a text, a program based on the presented approach must attempt to recognise the G.G. of the word by checking first the last 7 characters of the word to see if they match any of the members of the G.G. database. Seven characters is the maximum length of a G.G. If this fails, it must start an elimination process, eliminating one character at a time (always left-right) and checking every new G.G. against the G.G. database. The program will always come up with a matching G.G. The inclusion of G.G. consisting of one character, i.e. *-α* or *-η* aimed primarily at catering for those noun endings that are not frequent enough to count as typical G.G. of Modern Greek and therefore to be explicitly included in our corpus.

An opposite direction process (right-left) must start the recognition process from the last character of the entry. Then it will immediately come up with an answer, thus considerably reducing matching times. However, a one or two-letter G.G. is very likely to belong to the non-specifying section, consequently providing two to three alternatives. Thus, time saved on the matching process will inevitably be spent on the recognition process.

(2) **Noun assessment**

Once the ending of the noun has been recognised in the form of a G.G. in the database, the program must start the assessment process. If the G.G. belongs to the specifying section, the noun will be assessed in terms of the G.N.C. pattern declared by the G.G. If the G.G. belongs to the non-specifying section, the noun will be assessed in the same terms but with a number of options, those offered by the recognised G.G. If, for example, the recognised G.G. is *-τίας,* the noun can be either masculine/feminine gender, singular number, nominative case, i.e. *κτημα-τίας;* or feminine gender, singular number and genitive case, i.e. *λαοκρα-τίας*. In both cases the program must first check the exception list that appears under the corresponding G.G.

(iii) **Base-form rules**

At the final stage the program must provide instructions for the lemmatisation modifications. If the lexical item is already in singular nominative form, it must be identified as such and no further action must be taken. If it occurs in any of its other inflected forms, the program must follow certain rules - provided under the corresponding G.G. - and execute the lemmatisation command. No problems are anticipated when the G.G. is non-specifying. The program must simply check in every alternative the singular nominative reporter. If this is not found, the lemmatisation rules must be implemented and the various possible Base-forms of the noun must be formulated. Only at this stage there is a need for either dictionary consultation or a certain interactivity module with the user who will be the one to define the existing Base-Form.

The actual program produced in 1992 by two Greek physicists at the University of Athens (G. Halkiadakis, G.Tsiatouhas) opted for the second approach (right-left) as regards the direction followed for the recognition of the G.G..They produced almost 1000 rules on the basis of our database which took up very little space in the computer's memory - less than 140 kbytes - and although they used PROLOG as the programming language, they were of the opinion that any programming language could be employed for such task. They tested their program against 150,000 entries (mainly nouns, adjectives and participles) in all their forms and they concluded that the horizontal approach proposed hereby for the automatic assessment of nouns in terms of their Gender/Number/Case pattern without any dictionary consultation was 99,1% successful.

PROGRAM GENERAL APPLICATIONS

Given the small volume of the database and the fast processing capacity of today's technology, the proposed method can be incorporated in a number of computerised tools, such as:

a. educational programs aiming at teaching Modern Greek grammar to the first classes of Primary school.

b. Greek text processing programs for a quick control of the spelling of words. Functioning as a preliminary filter, the lemmatisation rules can automatically provide the Base-form of a word which in turn can be checked against a stem-only dictionary.

c. Greek computerised dictionaries for teaching Greek as a foreign language. The lemmatisation rules will automatically provide the learner with the Base-Forms of his/her unknown words.

d. Machine Translation programs for the automatic lemmatisation and recognition of the grammatical category of Greek words in context.

REMAINING TYPES OF AMBIGUITY

Once the lemmatiser has come across specifying G.G., lexical items in a sentence can be immediately recognised. However, what has become obvious so far is the fact that there will always be a number of ambiguous cases which even the non-specifying morphological speculations cannot resolve. In the sentences

Με χτύπησε απαλά στην πλάτη
he-me patted softly on the back
Πόσο μου αρέσουν τα απαλά χέρια σου
how I-love (the) soft hands yours

the word *απαλά* (soft-ly) can be given a definite meaning only after its grammatical category (an adverb or an adjective) has been established syntactically. It is a case similar to the word *hard* in English.

Moreover, even if the lemmatiser is able to help in homograph resolution before the syntactic analysis of a sentence, there are still extreme cases in which the ambiguty can only be resolved by contextual analysis. The presence of articles and adjectives in a nominal phrase, which agree in terms of gender, number and case with the head-noun of the NP, does not always reveal the specific grammatical and lexical meaning of this noun. In a sentence for instance such as

Τρελλαίνομαι στη θέα ωραίων κερασιών
I get crazy at the view of beautiful cherries

it is not clear whether we go crazy over beautiful cherry tress or a bowl of fresh dark red cherries to devour. Both nouns, *κεράσια* (cherries Nt., P., N.A.V.) and *κερασιά* (cherry tree, F., S., N.A.V.) formulate the genitive plural case by employing the same inflectional suffix -*ιων,* the accent being on the last syllable. The adjective can be of no help either, as it can refer to both fruits and tree, both morphologically and syntactically. The answer lies in the context, either linguistic or pragmatic.

CONCLUSION

The findings were beyond initial expectation. Six hundred and fifty one **specifying** endings for the nominal group - out of a total of eight hundred and fifty -, is a promising figure. Research continues in order to establish all the specifying G.G. of the two other large declinable Greek grammatical categories: adjectives and verbs. As regards pronouns and articles, there are 220 inflecting forms which can be coded as unambiguous in the dictionary. Undoubtedly, interesting conclusions can be drawn on the morphological system of the Modern Greek language. Moreover, given the fact that grammatical disambiguation by means of morphology has always been desirable in the areas of Machine Translation and spelling checkers, it is hoped that the proposed approach will contribute in this direction.

BIBLIOGRAPHIC NOTES

(1) Babiniotis G., Theoretical Linguistics: Introduction to Modern Linguistics, pp. 164
(2) Quirk R. and Greenbaum J., A University Grammar of English, pp. 18-19
(3) Triantafillidis M., Modern Greek Grammar, pp. 66

SELECTED BIBLIOGRAPHY

1. Γεωργοπαπαδάκος, Α., Το Μεγάλο Λεξικό της Νεοελληνικής Γλώσσας (Μονοτονικό), εκδ. Μαλλιάρης, Αθήνα, 1984
2. Ιωαννίδης, Κ., Νέο Λεξικό της Γλώσσας μας, εκδ. Αιγαίο, Θεσσαλονίκη, 1985
3. Κεσίσογλου, Ι.Ι., Το νεοελληνικό κλιτικό σύστημα, Ελληνικά 18, 1964
4. Κουρμούλης, Γ.Ι., Αντίστροφον Λεξικόν της Νέας Ελληνικής, Αθήνα, 1967
5. Κουχτσόγλου, Ν. & Γεωργακόπουλος Σ., Σύγχρονον Λεξικόν της Ελληνικής Γλώσσας, εκδ. Άτλας, Αθήνα, 1961
6. Λεξικόν της Ελληνικής Γλώσσης, εκδ. Πρωία, Αθήνα, 1933
7. Μηχιώτης, Χ., Νεώτατον Λεξικόν της Νεοελληνικής Γλώσσης (καθαρευούσης-δημοτικής), εκδ. Κασταλία, Αθήνα, 1972
8. Μπαμπινώτης, Γ., Θεωρητική Γλωσσολογία: Εισαγωγή στη Σύγχρονη Γλωσσολογία, Αθήνα, 1980
9. Πρότυπον Λεξικόν της Νέας Ελληνικής, Β' έκδοση, εκδ. Σταφυλίδη, Αθήνα, 1952
10. Σταματάκος, Ι., Λεξικόν της Νέας Ελληνικής Γλώσσας, εκδ. Δημητράκου, Αθήνα, 1952
11. Τριανταφυλλίδης, Μ., Μικρή Νεοελληνική Γραμματική, εκδ. ΟΕΔΒ, Αθήνα, 1949
12. Τσιατούχας, Γ. & Χαλκιαδάκης, Γ., Αυτόματη Αναγνώριση Παρεπόμενων Ουσιαστικού σε PROLOG, διπλωματική εργασία, Πανεπιστήμιο Αθηνών, Τμήμα Φυσικό, Αθήνα,1990

13. Τσολάκης, Χ., Νεοελληνική Γραμματική της Ε' και ΣΤ' Δημοτικού, εκδ. ΟΕΔΒ, Αθήνα, 1981
14. Τσοπανάκης, Α.Γ. & Δερβισοπούλου, Μ., Το κλιτικό μας σύστημα, Επιστημ. Επετηρίς της Φιλοσοφικής Σχολής του Πανεπιστημίου Θεσ/νίκης, τομ. 7, 1956
15. Φλώρου, Αθ., Νεοελληνικό, Ετυμολογικό και Ερμηνευτικό Λεξικό, εκδ. Λιβάνη, Αθήνα, 1980
16. Caradec, R. & Saada G., Definition de la Classe Syntaxique d' une Forme Lexicale a partir de sa Terminaison Graphique, Linguisticae Investigationes, TOME VI:2, 271-281 John Benjamins B.V., Amsterdam, 1982
17. Hartman, R.R.K., Lexicography: Principles and Practices, Academic Press, London, 1983
18. Lehnert, W.G. & Ringle, M.H., Strategies for Natural Processing, Lawrence Ertbaum Associates, 1982
19. Mackridge, P., The Modern Greek language: A Descriptive Analysis of Standard Modern Greek, Oxford University Press, N.York, 1985
20. Quirk, R. & Greenbaum, S., A University Grammar of English, Longman, London, 1976
21. OXEYE: a Text Processing Package for 1906 A, Oxford University Computing Laboratory, 1976
22. SYSTRAN CODING MANUAL, TELINDUS S.A., Luxembourg, 1990