Semantic Modeling in Morpheme-based Lexica for Greek

M. Grigoriadou, E. Papakitsos & G. Philokyprou
University of Athens, Faculty of Science, Dept. of Informatics, Section of Computer Systems and Applications, Panepistimiopolis, TYPA Buildings, 157 71 ATHENS, Greece.
e-mails : gregor@di.uoa.gr, papakitsev@vip.gr.

Abstract
A Machine Readable Dictionary (MRD or Lexicon) can be designed as a large-scale lexical database, having the task of supporting many different applications such as morphological, syntactic and semantic processing, information retrieval, machine translation, educational tools, etc. Regardless of how different these applications may be, they need a comprehensive lexical database to rely on, since it is quite wasteful to develop a different lexicon for each application. This paper deals with a method for designing and organizing a multi-purpose morpheme-based lexical database for Greek. The authors are in favor of morpheme-based lexical databases in order to avoid a repetition of effort from one application to another, and in order to achieve flexibility, reusability and expandability. The proposed method for modeling the lexical database is the Entity/Relationship model, which was originally designed for natural language processing. Even though our system was tested in Greek language, these methods could also be applicable to other languages having similar morphological systems to that of Greek.

Key words:
artificial intelligence, lexical database, semantic modeling.

1. Introduction

In the work presented here, a lexical database was originally developed to support a morphological processor for Modern Greek. The morphological processor is a modified version of M.I.T.-Decomp [Allen, Hunnicutt and Klatt (1987), Sproat (1992)], based on functional decomposition and adapted for Modern Greek [Papakitsos, Gregoriadou, Ralli (1998)]. The design of the lexicon is based on the approach of how processors (taggers) are simplified by the use of a large scale lexicon rich in linguistic information (regular and idiosyncratic). For every lemma there is information about morphosyntactic and semantic relations, derivation, compounding, pronounciation and others. This approach can satisfy better the long term criteria of expandability, reuseability and simplicity, since the enriched lexical database can also support semantics, generation and code minimization [Papakitsos et al. (1998)].

The lexical database and the entire project aims on researching and developing language engineering tools for Modern Greek, following similar work presented for other languages [Dura (1994), Goñi, Gonzalez and Moreno (1997), Mikheev, Liubushkina (1995)]. Until today, many attempts have been made to develop MRDs in Modern Greek. A variety of lexical databases were developed in order to support specific applications or to test theoretical models. Most of the above are mainly dealing with inflection . Exceptions to this are:

(a) a system that was used in the EUROTRA project [Ananiadou, Ralli, Villalva (1990)],
(b) a lexicon for supporting a commercial system of spell-checking [Karalis (1993)] and
(c) the processor ATHINA, which was developed to test the applicability of Generative Lexical Morphology to Modern Greek [Ralli (1985), Ralli & Galiotou (1987), (1991)].

Sgarbas, Fakotakis and Kokkinakis [(1995)] and Markopoulos [(1997)] developed MRD systems supporting the two-level morphology model [Karttunen (1983), Koskenniemi (1983)]. Before proceeding to the design of the lexical database, a thorough study of the domain (Modern Greek morphology) was imperative in order to isolate its characteristics. A brief description of Modern Greek follows along with the associated aspects of its morphology using the theory of Generative Lexical Morphology, as it was adapted for Greek by Ralli [(1983),(1985),(1986),(1988),(1992a,b), (1994)].

2. Greek Morphology

Greek is a language of concatenative morphology, where morphemes constitute the basic units of morphological processes. The three major morphological processes are inflection, derivation and compounding.

Greek has a rich nominal inflectional morphology like Slavic languages or Latin. In a dictionary of average size, containing 60,000 entries, less than 1,700 words can be found without an ending. An ending is generally added to a bound morpheme (stem) in order to form a word. Nominals have four cases and two numbers. Nominal or adjectival endings are characterized accordingly (eg. ημέρ-α *day*, gender: feminine, number: singular, case: nominative/accusative/vocative). Stress is orthographically marked in Greek; It has a phonological function and it plays a very important role in the morphological and phonological language systems (compare πότε = *when* and ποτέ = *never)*. Verbal morphology is more complex. A verbal paradigm can have more than 50 words per paradigm, where endings are marked for person, number, aspect, tense and voice. Computationally, more than 60 inflectional classes (i.e., inflectional paradigms) can be identified, but the accurate number of them depends on the approach, (though linguistically, it has been claimed that these categories are much less than 60-cf. [Ralli (1994)]).

Derivation is also quite productive. More than 58% of the words found in a dictionary of average size are derivatives. Generally, the addition of a suffix may change the stem category, while most of the prefixes do not change the category:

Suffixation: δώρ-ο 'the gift' and δωρ-ίζ-ω 'to make a gift'

Prefixation: γράφ -ω *to write* and ανα- γράφ -ω *to inscribe.*

As in German or Swedish, compounding is an important part of Greek morphology, and it is traditionally defined as an association of two or more stems which always form a single unit. Compounds are generally classified as nouns, verbs and adjectives. They make up more than 29% of the entries in a dictionary. In most cases a linking vowel "ο" connects the two stems (εθν-ο-φρουρά : national guard. For more details on Compounding in Greek, see [Ralli (1992a)].

According to our research, Modern Greek contains approximately 8000 morphemes of the following classes: 1700 free-morphemes, 5900 roots, 150 endings and about 200 prefixes and suffixes [Papakitsos et al. (1998)]. In order to get these 8000 morphemes isolated, two well-known average size dictionaries of Modern Greek [Papyros (1973), Tegopoulos-Fytrakis (1993)] were analyzed, word by word. This process lasted more than a year, and it was carried out without the help of a computer. From the previous description, it is clear that Greek is heavily depended on a small number of morphemes for the word production being done through the morphological processes of inflection, derivation and compounding. Despite the small number of morphemes, even lexica which contain up to 90000 stem-entries (see [Vagelatos, Triantopoulou, Tsalidis, Atmatzidi, Christodoulakis (1994)]) can encounter difficulties in supporting spell-checking. This happens because many novel words are not recognized, such as "αυτοεκτίμηση" (self-respect) or " ενδοδίκτυο " (intranet), although non of their constituent morphemes are novel ones [Papakitsos, Gregoriadou (1999)]. Consequently the development of a lexical database that supports full-scale morphological processing (and not only inflectional morphology) is useful.

3. Database Modeling

Having the characteristics of the language (i.e. Modern Greek), the main points of the initial process were:
(i) The selection of data-features to be encoded
(ii) To find methods of encoding these features
(iii) The efficient organization of the encoded features.
The data to be encoded were only selected to support the tagger, namely the morphemes (affixes, roots/stems, free-morphemes), their features (mainly morphosyntactic and stress assignment) and their relations (morpholexical & conversion rules, inflection, derivation and compounding). The selection of morphemes is not a trivial task, (especially for the roots/stems) mainly because of the "portmanteau-morpheme" phenomenon and of the various versions of Greek. The selection of morphemes is a linguistic task that eventually does not affect the design of the data structures of the lexicon.
The examined database models for designing the database are the relational, semantic (entity/relationship) and object-oriented (for more details: [Date (1990), (1995)]). The relational model is supported by all major database developing systems. One of the main characteristics of this model is the easy data management. The entity/relationship model (E/R-model) was originally designed for the needs of natural language processing. It can describe the relations between the various items (morphemes) of the database and additionally it can considerably facilitate the task of a tagger by providing fast access to lemmata and to their features. In a lexical database, data-management and tagger-support are not equally important functions. For this reason, the object-oriented and the E/R models were considered as supplementary rather than competitive. The object-oriented model was initially used for designing the data management system of the lexicon, comprising of six files and having a user-friendly environment. Each of the first four files contained one class of morphemes (prefixes, suffixes, free-morphemes and roots/stems) with their associated features. The fifth file contained the endings with their features arranged in paradigms, and the sixth file contained the derivatives , the compounds and their features. The object-oriented model was used in order to avoid a great deal of redundancy, i.e. thousands of nouns, verbs, adjectives and adverbs share the same attributes which are constantly repeated. Additionally, the attributes of endings can be allocated only when the specific paradigm is known, since an ending may belong to more than 4 paradigms (e.g. -α can be attached to verbs, nouns, adverbs, adjectives and pronouns!). Thus, the design was shifted towards an object-oriented direction. Classes and subclasses were defined (i.e. verbs, nouns, etc. and their paradigms) and each lemma inherits its attributes according to the class that belongs to (see [Gazdar, Kilbury (1994)]). The controversy about object-oriented modeling compared to relational modeling in data management systems [Date (1990)] was examined and it was eventually decided that since the lexicon is not a general purpose database, but one tailored to this specific domain (NLP), the controversy could be overlooked, particularly because object-oriented modeling seemed to work well for the application in hand. It must, however, be noted that the relational and object-oriented model can be interchangeable for the first part of the process, if it is decided to use a standard database management system tool to support an application. In that case, object-oriented techniques can be used for the designing of a user-friendly interface including forms or reports [Sjögreen (1994), Bekos (1998)].

There is a great deal of work to be done when a lexical database is initially enriched with data but once this stage is over, the relevant modeling can be discarded. The database modeling is an internal process of the system hidden from the user thanks to the interface environment. After the full construction of the lexicon, a more appropriate model can be used to organize the lexicon in order to facilitate morphological processing.
This model is the E/R model, which was originally designed for NLP needs [Date (1990)]. Each term of E/R modeling has an equivalent linguistic term. The rules of the lexicon are designed as relations among entries (stems, derivational and inflectional affixes). The entries and their relations between them are diagrammatically depicted, and they can be implemented

in a quite straightforward way. In Fig. 1, for every morpheme class (rectangle) there is a description of its relations. The "is-a" relations of the allomorphs are depicted by a rhombus. The double-head arrows denote a "one-to-many" relationship, i.e. that a morpheme may have more than one allomorphs (e.g. the case of some Greek verbs). All types of "has"-relations are depicted as arrows. A root (on top) may be related to particular prefixes (prefixation), suffixes (suffixation), endings (inflection) or other roots (compounding). Suffixes can be related to others or to endings (inflection). Some words (free-morhemes) can be related to a prefix (adverbs like "παρα-έξω": farther out). Finally, a stem can be related to a root, because of derivational or compounding processes ("Origin").

In the above way, all of the words having a common root form a tree called the **family tree**. The root of the family tree is the common root (morpheme), which is connected to all the affixes that can be combined with it. An example of such a family tree is the following one [Ralli (1985)]. The words:

γράφ-ω	(to write)
υπο-γράφ-ω	(to sign)
ανα-γράφ-ω	(to inscribe)
κατα-γράφ-ω	(to record)
δια-γράφ-ω	(to delete)
εγ-γράφ-ω	(to register)
περι-γράφ-ω	(to describe)
παρα-γράφ-ω	(to erase in law)

have a common root ("γραφ-"). This common root is the root of the family tree. The left branch (nodes) of this family tree is an array containing the above prefixes (υπο-, ανα-, κατα-, etc.). The right branch of the root may contain suffixes that can be attached to the common root, like "-ικ-" (as in γραφ-ικ-ός: writing, graphic, picturesque, …) or "-ει-" (as in γραφ-εί-ο: desk, bureau, office). In a similar way, the suffix "-ικ-" (it produces adjectives) can be connected to each one of the nodes of the left branch (on their right as well) to produce words like "περι-γραφ-ικ-ός" (descriptive). Following that, the suffix "-ικ-", as an entry of its own, can be connected in a similar way to the suffix "-οτερ-". From there, the recursive mechanism of our system can recognize words like "γραφ-ικ-ότερ-ος" (more picturesque) or "περι-γραφ-ικ-ότερ-ος" (more descriptive), without the need of storing them in the lexical database.

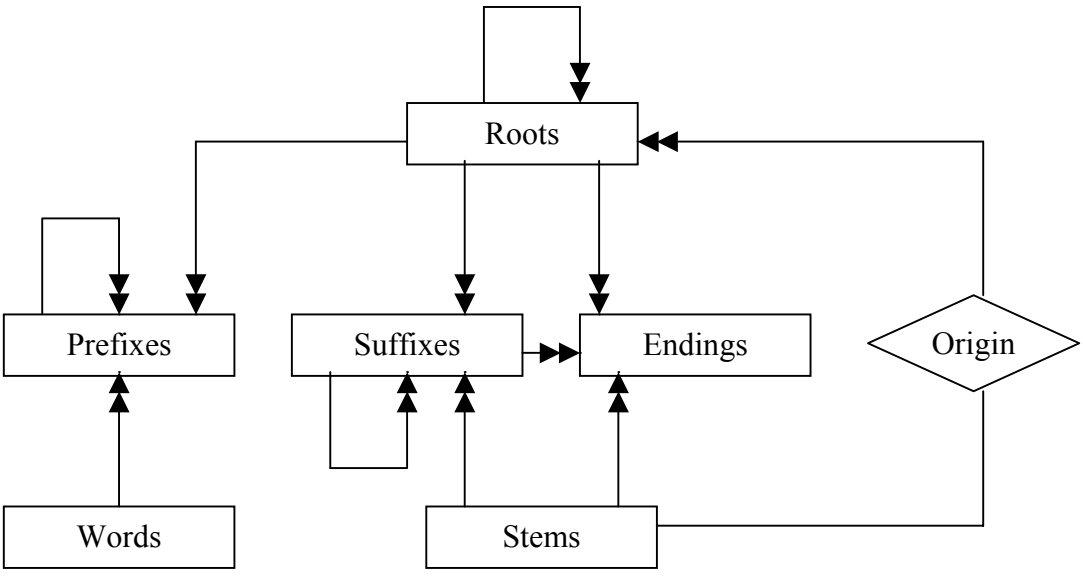

Figure 1. The "has"-relations between the entities of the database.

The explicit expression of these relations is necessary in Greek because otherwise taggers (having a connection of morphemes through morphemic features, as in SIL's AMPLE, see

[Sproat (1992)], as in DATR or as in other similar approaches) will not be able to deal with them effectively. For example, there are seven suffixes in Modern Greek that can be attached to a nominal stem in order to derive a verbal stem (i.e. –ιζ-, -αιν-, -ων-, -ευ-, -ιαζ-, -αζ-, -αρ-), without any apparent criterion existing for the six of them (e.g. to the root δωρ- {root of the word gift} only the suffix –ιζ- can be attached, to produce a verbal stem {δωριζ-}, see [Papakitsos et al. (1998)]). This characteristic does not affect so much the analysis mode of the processor, but it does considerably affect the generation mode by producing a large number of non-existing stems (e.g. δωρευ-, δωρων-, δωραζ-, etc.). All of the seven noun-to-verb suffixes have the same categorial features and consequently overgeneration can not be prevented. In the present system's way of connecting morphemes (explicitly), the overgeneration of stems is avoided and semantic relations can be enforced more easily as well (through the family trees). The previous description (of Fig. 1) is based on the linguistic theory of Generative Lexical Morphology, as it was adapted to Modern Greek [Ralli (1983), (1986)]. The designing of the lexical database according to E/R-modeling allows all the characteristics of that linguistic theory to be enforced in the lexical database, through the "is-a" or the "has" relations.

4. Implementation

The entries of the lexicon are implemented as variables or records (groups of them) and their relations (the arrows of Fig. 1) are implemented as pointers denoting either a record number or a displacement from one record to another. The result of the implementation is a set of files logically connected together as shown in Figure 2. According to this implementation, a file ("Size") contains the maximum number of characters for every type of morpheme. Every entry is directly associated with certain fields (1 through 11 in the figure). The "Class" field denotes the type of morpheme (prefix, suffix, root, etc.). The "String"-field is the orthographic or phonetic description.
The "Next"-field is a pointer to another entry of the same form but having different features (e.g. "man" the noun and "man" the verb). The "key"-field is a unique number which is allocated to every entry (lemma). The allocation of keys is a standard process in databases that ensures future expansion without unnecessary repetition of data, it also provides unique identification. Compared to a string, the key-number is a more compact representation of the lemma, having fixed-size (either 2 or 3 bytes long instead of an average length of 6.5 bytes per string). Moreover, it is used to identify uniquely entries of the same form but of different attributes (e.g. "man" the noun and "man" the verb would be mapped to different key-numbers). Other relations include allomorphs and inflectional paradigms. All the allomorphs of a stem are connected together through an integer denoting the record number of an array ("Index of Allomorphs"), where all the key-numbers of the allomorphs are listed together ("List of Allomorphs"), as in the following example:
{5126, "μυλων-άς", 896: miller}
{3871, "μυλωνάδ-ες", 1503: millers}
{"File of Allomorphs": Record 896 = {5126, 3871}}.

The inflectional paradigm of a stem is designed in a similar way to the above one. This is an integer ("Index of Paradigms") denoting the record-number of an array where all the key-numbers of the endings, that belong to this paradigm, are listed together ("List of Endings"). Prefixation, suffixation and compounding are designed in a way similar to the inflectional paradigm or similar to the allomorphs' relations. In suffixation for example, the stems are connected to a file, where all the key-numbers of the suffixes, that can be directly attached to their right, are listed there. The attributes of an entry (the attributes are 2 to 4 bytes long) are connected to the relevant entry through an integer (byte) denoting the record number of the attributes' string in the associated array. The attributes include category (part-of-speech), subcategories, case, number and gender for nominals, stress assignment for free-morphemes and endings, and others. In the above way, by using sparse matrix encoding schemes

[Tremblay, Sorenson (1984)], the access to attributes and their related entries is direct, and the memory requirements are relatively low. Additionally, long-distance dependencies can be dealt with a couple of techniques. One of them is to connect the prefixes of a root to the relevant suffixes through pointers. The other is to map the combination of a root and its associated affixes to a new entry and then to treat the new entry accordingly.

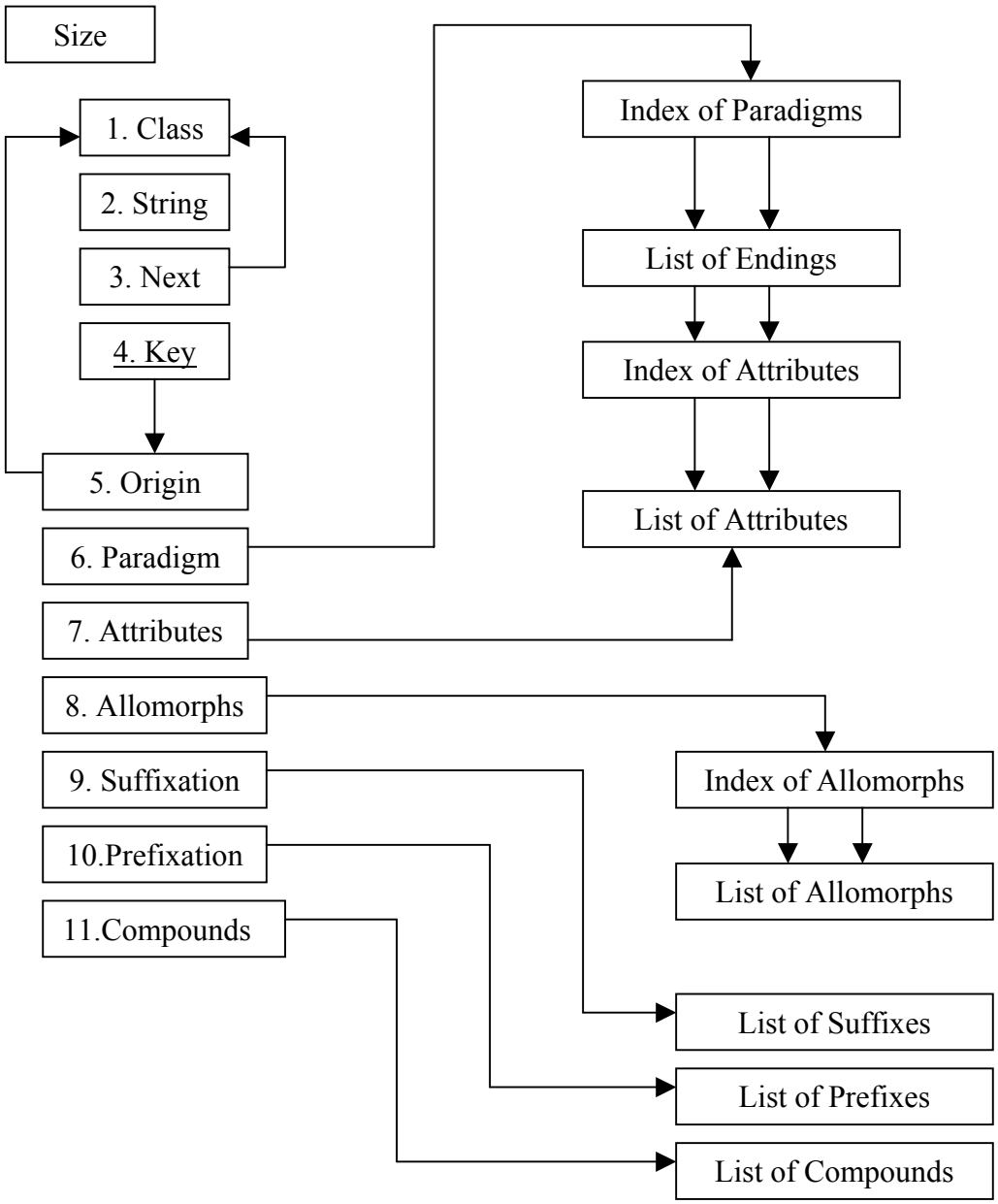

Figure 2. The functional diagram of the lexical database.

5. Results

Our system was tested on the Greek part of the ECI (European Corpus Initiative) which is a large scale corpus, a joint project of the Universities of Edinburgh and Geneva for ACL. This corpus, containing over 1,879,000 words, is actually composed of only 88,974 different words, where a spelling-error rate of approximately 2% was observed. These different words are produced by 32,629 lexemes, consisting of 1669 free morphemes, 1542 root-Infl.Affix lexemes and 25,202 derivatives and compounds. From the above figures, approximately

7800 entries (1669 free-morphemes, 5758 roots, 149 endings and about 200 prefixes and suffixes) were initially extracted to make our morpheme-based lexicon. This lexicon was constructed semi-automatically, i.e. with the help of supporting tools being developed for this specific purpose. The construction process was implemented in the following steps:

- The word-tokens of the corpus were automatically isolated, along with their frequency of appearance in the text.
- The word-tokens were separated (semi-automatically) according to their internal structure (i.e. free-morphemes, root-ending words, derivatives and compounds were gathered in different files).
- The endings were stripped off from the word-tokens in two stages: The first stage identified the ending (automatically) and the second was validating the first stage (semi-automatically).
- The discovered endings and roots were separated in different files.
- The previous two steps were followed for finding suffixes and prefixes (as it happened with endings).
- Some of the morphemes were marked with their linguistic features, for testing purposes.

In the above process, every word is classified according to its internal structure. There are twenty such classes for derivatives (the compounds were not studied thoroughly). For example, the word "περι-γραφ-ικ-ός" (descriptive) has the internal structure [prefix-root-suffix-ending], the word "γραφ-ικ-ότερ-ος" (more picturesque) has the internal structure [root-suffix-suffix-ending] and the word "γραφ-εί-ο" (desk, bureau, office) has the internal structure [root-suffix-ending]. These three words, although have a different internal structure, share a common root ("γραφ-"). Thus, they belong to the same family tree (see section 3. Database Modeling). By constructing the family trees (semi-automatically), all the morphological relations between the morphemes in the database are automatically identified.

6. Conclusion

Our objective was to evaluate the performance of our lexical database regarding the following points:
(a) to cover inflection, derivation, compounding and long-distance dependencies,
(b) to improve simplicity both of usage and of the design.
(c) to examine how the size of the lexicon affects the above.
It was demonstrated that our tagger can produce robust results (98.2% accurate tagging on the Greek part of the ECI), provided that it is supported by a morpheme-based lexicon enriched with idiosyncratic information maintaining a high recognition speed of more than 1100 words/sec [Papakitsos et al. (1998)]. Concerning the results of the tagger, an unspecified number of words are given two analyses, both of them correct, considering the abilities of the tagger. These are words like "αναθέσεις", which can be a noun (= assignments) or a verb (= to assign: imperative, 2nd person). The disambiguation can be done only at a syntactic level, since each case is preceded by different words (conjunctions or articles). Towards that direction, it was also demonstrated that the database modeling can offer the required results by developing a larger and more reliable lexical database. According to small-scale data tests, it is believed that the processing error will be decreased to about 0.5% in future, by using rich linguistic information, being encoded as attributes for every lexical entry.

Acknowledgements
Thanks go to As. Prof. A. Ralli for her contribution in the linguistic part of this research.

References

Allen, J., Hunnicutt, M.S. and Klatt, D. (1987*). From Text to Speach: The MITalk System.* Cambridge University Press.

Ananiadou, S., Ralli, A., Villalva, A. (1990). The Treatment of Derivational Morphology in a Multilingual Transfer-Based MT System-EUROTRA. In Proceedings of *SICONLP '90*, Seoul.

Bekos, E. (1998) Implementation of an interface system and of a preprocessor for supporting a morphological processor of Modern Greek, Diss. Reg. No 329, Dpt of Informatics, University of Athens [in Greek].

Date, C.J. (1990). *An Introduction to Database Systems*. 579-610. Volume I, Fifth Edition, Addison-Wesley.

Date, C.J. (1995). *An Introduction to Database Systems*. Volume I, Sixth Edition, Addison-Wesley.

Dura, E. (1994) Lexicon and Lazy Word Parsing. Proceedings of the Language Engineering on the Information Highway Conference, ILSP, Athens.

Gazdar, G., Kilbury, J. (1994). Lexical Knowledge Representation. Course CA1, ESSLLI'94, Copenhagen Business School.

Goñi, J., Gonzalez, J. and Moreno, A. (1997). ARIES: A lexical platform for engineering Spanish processing tools. *Natural Language Engineering*, Vol. 3(4), Cambridge University Press.

Karalis, K. (1993). For correct Greek. *RAM-February,* Athens [in Greek].

Karttunen, L. (1983). KIMMO: A General morphological processor. *Texas Linguistic Forum*, 22:165-186.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.* Ph.D. thesis, University of Helsinki.

Markopoulos, G. (1997). A two-level description of the Greek noun morphology with a unification-based word grammar. *Working Papers in NLP*, 175-188. Diaulos Public., Athens.

Mikheev, A., Liubushkina, L. (1995). Russian morphology: An engineering approach. *Natural Language Engineering*, Vol. 1(3), Cambridge University Press.

Papakitsos, E., Gregoriadou, M., Ralli, A. (1998). Lazy Tagging with Functional Decomposition and Matrix Lexica: an Implementation in Modern Greek. *Literary and Linguistic Computing,* 169-176. Vol. 13(4), Oxford University Press.

Papakitsos, E., Gregoriadou, M. (1999). Matrix Lexica: An alternative description of lexical databases, In proceedings of the 2nd Conference on Hellenic Language and Terminology, ELETO, Athens.

Papyros Publ. (1973). *Orthographical Dictionary*, by N. Sifakis. Athens [in Greek].

Ralli, A. (1983). Morphologie Verbale et la Theorie du Lexique: Quelques Remarques Preliminaires. In proceedings of the 4th Annual Meeting of the Linguistic Section, Univ. of Thessaloniki [in Greek].

Ralli, A. (1985). *Morphology*. Prep.Phase,Volume I, Chapter 2, Eurotra-Gr, Athens.

Ralli, A. (1986). Inflection and Derivation. In proceedings of the 7th Annual Meeting of the Linguistic Section, Univ. of Thessaloniki [in Greek].

Ralli, A. (1988). *Elements de la morphologie du grec moderne: la structure du verbe*. PhD diss., Universite de Montreal.

Ralli, A. (1992a). Compounds in Modern Greek. *Rivista di Linguistica*, Special issue on Compounds, Vol. 4(1): 143-174. Scuola Normale Superiore, Pisa.

Ralli, A. (1992b). The theory of features and the structure of the inflected words of Modern Greek. In proceedings of the 13th Annual Meeting of the Linguistic Section, Univ. of Thessaloniki [in Greek].

Ralli, A. (1994). Feature representations and feature-passing operations: the case of Greek inflection. In Proceedings of the *8th Linguistic Meeting on English and Greek*, Univ. of Thessaloniki.

Ralli, A., Galiotou, E. (1987). A Morphological Processor for Modern Greek. In Proceedings of the *3rd European ACL Meeting*, Copenhagen.

Sgarbas, K., Fakotakis, N., Kokkinakis, G. (1995). A PC-KIMMO-Based Morphological Description of Modern Greek. *Literary and Linguistic Computing*, Vol.10(3): 352. Oxford Univ.Press.
Sjögreen, C. (1994). *Descriptions to some of the GLDB frames*, Dept. of Swedish Language, Goteborg University.
Sproat, R.W. (1992). *Morphology and Computation*, MIT, USA.
Tegopoulos-Fytrakis (1993). *Greek Dictionary*, Armonia Publ. Athens [in Greek].
Tremblay, J., Sorenson, P. (1984). *An Introduction to Data Structures with Applications*. McGraw-Hill.
Vagelatos, A., Triantopoulou, T., Tsalidis, C., Atmatzidi, M., Christodoulakis, D. (1994). Correcting Spelling Errors in Modern Greek by use of a Lexicon. Proceedings of the Language Engineering on the Information Highway Conference, ILSP, Athens.