Literary and Linguistic Computing Advance Access published February 20, 2012

Design and implementation of an electronic lexicon for Modern Greek

Panagiotis Gakis, Christos Panagiotakopoulos and Kyriakos Sgarbas
University of Patras

Christos Tsalidis
"Neurolingo" Company

Abstract

Natural language text analysis presupposes the encoding of morphological phenomena. In this article, we present some particularities of Modern Greek and the way these are encoded in the presented electronic lexicon. The project plan of its development combined both simple planning algorithms and more elaborate ones for the generation and recognition processes. The resulted lexicon exhibits fast access to its contents and easy content management. It is re-usable and modular enough to support existing NLP applications.

Correspondence:
Department of Primary Education, University of Patras.
E-mail:
gakis@sch.gr

1 Introduction

Electronic lexicography (De Schryver, 2003; Carr, 1997; Perry, 1997) covers a wide range of tasks, including project planning, compilation, usage and finally evaluation of the electronic linguistic resources, with the help of computational tools. Electronic linguistic resources can be: (i) electronic corpora, (ii) electronic-computational lexicons for natural language processing (NLP) systems.

There are two categories of electronic lexicons:

- Machine readable dictionaries (MRDs), usually built by transcription of existing printed-form lexicons to computer-readable form, aiming to support larger NLP systems.
- Electronic lexicons for humans, which employ a user interface for the interaction and databases with extra information.

The basic advantage of electronic lexicons (in contrast to printed ones) is their ability to store arbitrary amounts of information in any of their fields: '*The absence of space constraints call for more, not less, intellectual discipline in the selection and arrangement of information*' (Hanks, 2001). Moreover they exhibit near-instantaneous response to recall, process, and exposition of data, capability of searching in many sub-lexicons simultaneously and ability for more frequent and timely updates than printed form lexicons.

An additional advantage of electronic lexicons is their augmented search functionality. Besides the traditional alphabetic search by lemma, it is possible to perform complex searches by using information in other fields as well, i.e. (i) grammatical and morphological information (e.g. all nouns with no plural form), (ii) style fields (e.g. colloquial words), (iii) thematic domain (e.g. biomedical terms), (iv) etymological information (e.g. words that have Ancient Greek origin), and (v) syntactic information (e.g. transitive verbs).

In this article, we describe the Neurolingo[1] Lexicon (http://www.neurolingo.gr/en/technology/lexica/morpholexicon.jsp), a general-purpose electronic lexicon for Modern Greek, of the MRD category (i.e. it is not appropriate for direct usage by

doi:10.1093/llc/fqs002

people, but instead it can support various computational applications such as morphosyntactic tagging, parsing, semantic tagging, machine translation, etc.) The Neurolingo (MRD) lexicon includes about 90,000 Modern Greek lemmas producing a total of 1,200,000 inflected word forms. The inflected word forms that are generated, are tagged using 67 parts-of-speech and morphological attributes, 79 domain and style attributes, 77 attribute sets, 19 accent rules, a 191 suffix rules, and 306 grammar rules.

The lexicon is descriptive and includes only the word forms used in contemporary oral and written speech.

2 Particularities of Modern Greek Language

Modern Greek grammar recognizes eleven parts of speech, six declinable, and five indeclinable. The declinable ones produce a huge set of morphologically inflected word forms, since Modern Greek is a highly inflectional language. For instance, from a single verb lemma more than 300 inflected word forms can be produced (including both active and passive voice word forms); from an adjective lemma about 100 word forms can be produced (if we include the comparative and superlative forms). Moreover, for many lemmas there are some additional archaic word forms that are still in use in colloquial Modern Greek.

Another complexity in written Modern Greek concerns the accents and the rules determining their position within the words. Spelling rules dictate that every word with two or more syllables requires an accent diacritic; omission of the diacritic is a spelling error. Therefore there is always an orthographic indication of the correct accent position. These positions usually change during declination, thus an additional effort is required by lemmatizers in order to find the correct correspondences between word types and lemmas.

During declination suffixes change according to the rules of grammar. In some cases stems change as well. The same lemma may have different morphological stems. For example in lemma [λέω = I say], we have five different stems [λεγ-], [λε-], [ειπ-], [π-], [λεχθ-] for the formation of the inflectional word forms for all modalities, voices, and tenses.

Lexical ambiguity is also an important phenomenon in Modern Greek. Lexical ambiguity occurs when a word type has more than one corresponding lexical entries (lemmas) or when the word is used with different meaning in figurative sense. Van Eijck (Van Eijck and Jaspars, 1996) defines the lexical ambiguity as information shortage for the word meaning. As a result we have to deal with a great number of ambiguous words, and unless their meaning is resolved by the context, this ambiguity may carry over to phrases or even whole sentences. Lyons describes the ambiguity that is noticed generally in language with the term lexical ambiguity (Lyons, 1977) and recognizes two different categories for it: (i) the homonymy, e.g. the pronoun [[σου] λέω = I tell you] and the possessive [το γλυκό [σου] = your candy] are homonymous and (ii) the multi-meanings ambiguity, e.g. the noun [πλάνη] has two meanings: plane (the tool), and delusion. Lexical ambiguity has direct relation to the reconstruction and set-up of lexicological entries (Boguraev and Pustejovsky, 1990) and almost always implies semantic ambiguity. Due to lexical ambiguity we can have:

(1) Similar word forms that correspond to the same lemma, e.g. the word form [κόρη = daughter] is the same in nominative, accusative, and vocative cases.
(2) Similar word forms that correspond to different lemmas with the same part-of-speech and morphological attributes, e.g. the word form [ματιών = of the eyes / of the looks] is genitive case in plural and may correspond to lemma [μάτι = eye], or to lemma [ματιά = look], with the same morphological attributes.
(3) Homographs, i.e. word forms that correspond to different lemmas with different part-of-speech and morphological attributes, e.g. the word form [απαντήσεις = answers (noun, plural)/you to answer (verb)] can be either second person, indicative case, future tense of the lemma [απαντάω = I answer (verb)],

or nominative, accusative, or vocative case—additional ambiguity—of the lemma [απάντηση = answer (noun)].

3 Other Modern Greek MRDs

During the past few years, computational lexicons for Modern Greek have been developed by Computer Technology Institute (CTI), the Institute for Language and Speech Processing, the Wire Communications Laboratory(WCL) at the University of Patras and the Software and Knowledge Engineering Laboratory (SKEL) at NCSR 'Demokritos'. The CTI Lexicon contains approximately 80,000 lemmas and is based on the CTI lexicon formalism for the description of inflected words. It has been used as the basis for the Greek spelling checker adopted by Microsoft for its word-processor MS Word. This lexicon has also been used recently for the development of the Word Net, a semantic network that includes for every lemma not only morphosyntactic but also semantic information. The lexicon of the Institute for Language and Speech Processing (ILSP) contains approximately 20,000 lemmas and has been developed in the context of the EC project LE-PAROLE, aiming at the natural language processing applications. The lexicon of the WCL contains approximately 35,000 lemmas and was exploited in the context of the EPET-II project MILTOS for the development of a fast morphological analyser. The SKEL lexicon consists of approximately 60,000 lemmas and has been developed in parallel with Ellogon, a general-purpose text-engineering platform that facilitates the development of new tools as well as their integration in different applications (Petasis *et al.*, 2003).

One of the most common platforms for morphological processing is PC-KIMMO, implementing the two-level morphological model by Kimmo Koskenniemi (Koskenniemi, 1983). The two-level model concerns the way of morphotactics disposition as well as the morphophonemic changes (Detorakis, 2009). The PC-KIMMO software (http://www.sil.org) facilitates the development of concurrent transducers to generate and recognize words. Sgarbas *et al.* (1995, 1999) have used PC-KIMMO to create a two-level model for the morphological description of Modern Greek covering nouns, adjectives, regular verbs, and participles. It uses thirty-six rules that handle diphthongs, control the position of the accent, verbal augment, and other specificities of Modern Greek.

Directed acyclic word graphs (Sgarbas *et al.*, 2000a; 2000b), machine learning techniques (Papageorgiou *et al.*, 2000; Petasis *et al.* 2000; 2001), as well as statistical methods (Tambouratzis and Carayiannis, 2001) constitute alternative models that have been developed for Modern Greek. Ntais (2006) has created a rule-based stemming algorithm especially for Modern Greek. This algorithm, however, has some limitations, since it ignores accent position. It can be augmented to include accents, but this would increase the complexity of the system (Detorakis, 2009).

Other lexical resources/applications developed for Modern Greek include: (i) a computational morphological and syntactic lexicon developed by the Institute for Language and Speech Processing (ILSP). The lexicon contains 66,000 coded lemmas with morphological and syntactic information, according to PAROLE[2] model, (ii) an independent electronic morphological lexicon of Modern Greek (Baldzis *et al.*, 2005a,b). (iii) an electronic lexicon for economic terms (Stock Exchange terminology) by the Department of Translation and Language Process (D.T.P.L.) of Aristotle University of Thessalonica (Tziafa, 2007).[3]

4 The Neurolingo Electronic Lexicon

The Neurolingo (MRD) electronic lexicon is based on a model specifically designed to confront the particularities of Modern Greek. It is a result of systematic work, at research level—in the areas of lexicography and NLP—as well as at the level of development of specialized electronic lexicons and computer systems for text checking and correction. The project plan of its development combined both simple planning algorithms and more elaborate ones for the generation and

recognition processes. The resulted lexicon exhibits fast access to its contents and easy update. The word types are analysed and produced by a morphological processor via a lexicographic editor (Fig. 4).

Within the morphological processor, the complete set of Modern Greek suffixes have been encoded, including additional information denoting the style of each corresponding word form (spoken, archaic, dialectical, etc.), and classified under distinct declension classes in separate parts of speech. This way the overall size of stored information was reduced, as well as the response times for recognition and generation of word forms. In addition, this way the lexicon can be easily updated since the only information that must be inserted for a new lemma is its declension and its stem.

4.1 Size and overview of the lexicon

The entries of the Neurolingo (MRD) lexicon contain the formal word forms of the lemmas with grammatical and morphosyntactic attributes encoded as label tags.

From the approximately 90,000 lemmas encoded in the lexicon about 1,200,000 inflected word forms are generated by the morphological processor. Each word form carries the following information: (i) spelling (the right spelling of each inflected word form); (ii) morpheme (the word form of the elements constituting each word form: prefix, stem, infix, suffix); (iii) morphosyntactic information (part of speech, gender, declension, person, etc.); (iv) style (spoken, archaic, dialectical, etc.); (v) terminology (additional tags denoting whether the particular word form is contained in any domain-specific sub-lexicons) (Fig. 1).

Figure 2 shows the number of lexicon word forms for each part-of-speech lemma and other statistics concerning ambiguity. The different—from a spelling point of view—word forms for each part of speech have been counted, e.g. the lexicon has 182,188 different—from spelling view—word forms of nouns, which means that if the word form is both a noun and a verb (e.g. [πρά-ξεις]<[πράξη = nominative, plural of lemma {πράξη}], [πράττω = 2nd person, singular, future of lemma {πράττω}], this has been counted not only as a noun but also as a verb. In the large

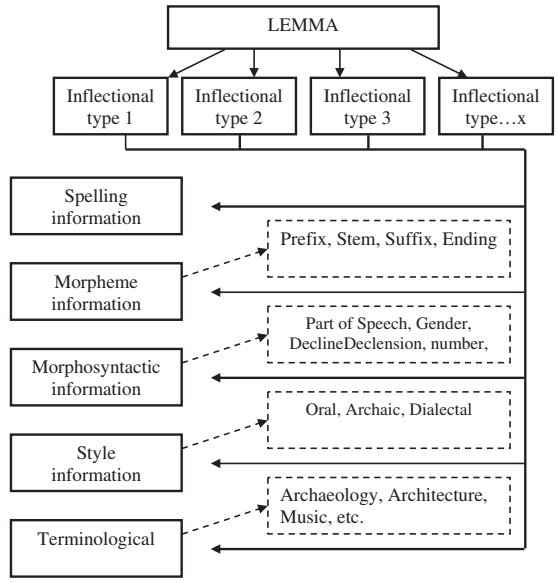

Fig. 1 Table of lemma's information

Part of Speech	Number of word forms	Lexical Ambiguity	Number of word forms
Nouns	**60,511**	Number of unique inflected word forms	**873,701**
Adjectives	**22,844**	Ambiguous word forms (from different Lemmas)	**39,119**
Verbs	**9,245**	Ambiguous word forms (from the same Lemma)	**4,758**
Participles	**865**	***Total number (for all ambiguous words)***	**43,877**
Adverbs	**7,830**		
Other parts of speech	**420**		
Total number (for all categories)	**101,715**		

Fig. 2 Statistics of the Neurolingo (MRD) lexicon.

number of ambiguous words from different lemmas 39,119 are included word forms that correspond to two or three lemmas, e.g. (φα-κών < [= genitive, plural of lemma {φακός = lens}, or genitive, plural of lemma {φακή = lentil}, genitive, plural

of lemma {φάκα = mousetrap}]. It is noteworthy that many word forms are repeated frequently during the declination of lemmas and 4,758 word forms differ only in their morphosyntactic attributes. For example the word form [διεθνή] in lemma [διεθνής = international] occurs with nine different sets of attributes: (1) genitive, singular, masculine, (2) accusative, singular, masculine, (3) vocative, singular, masculine, (4) genitive, singular, female, (5) accusative, singular, female, (6) vocative, singular, female, (7) nominative, plural, neuter, (8) accusative, plural, neuter, and (9) vocative, plural, neuter.

Apart from the general lemmas the lexicon also includes domain-specific sub-lexicons: currently one with 10,000 Greek toponyms (i.e. names of Greek regions, municipalities, districts, towns, villages, etc.) and one with 10,000 biomedical terms. More sub-lexicons are scheduled for development in the future.

The lexicon is available both as a platform-independent java library (.jar format) and in native form (C++ API for Windows, Unix, MacOSX). Its modular structure allows easy integration to many existing commercial applications: MS Word, Open Office Writer, Lotus Word Pro, Adobe InDesign, Quark Xpress, etc.

4.2 Sources of the lexicon

All lemmas in the Neurolingo (MRD) lexicon were gathered by indexing four of the biggest Dictionaries for Modern Greek: (i) Dictionary of Common Modern Greek Language by the Institute of Greek Studies of Aristotle University of Thessalonica (1998), (ii) Dictionary of Modern Greek by G. Babiniotis (1998), (iii) Greek Dictionary of the Modern Demotic Language by E. Kriaras (1995) and finally (iv) the Major Greek Dictionary Tegopoulos Fytrakis (1997).

Grammar rules, declensions, spelling rules, part-of-speech categories, and subcategories are all in accordance to Modern Greek grammar by Triantafyllidis (1998). In addition, especially for verb declination, we followed the structure of the more detailed book 'Verbs of Modern Greek' (Iordanidou, 1991). To ensure its completeness, the Neurolingo (MRD) lexicon is continuously and regularly updated by various sources (newspapers, WWW, etc.). It is currently in the 4th version after 8 years of development since its 1st version in October 2003.

4.3 Encoded information in word forms

For each lemma and for each generated word form, the lexicon provides the following information:

4.3.1 *Grammatical*

- on the lemma level its part of speech is encoded as one of the following eleven attributes: adjective (ADJ), adverb (ADV), article (ART), conjunction (CONJ), noun (N), participle (PART), particle (PARTICLE), preposition (PREP), pronoun (PRON), verb (V), and interjection (INTER).
- on the word form level there are additional attributes for: gender, case, voice, tense, person, degree, mood, declinable/indeclinable tag, and strong or slim word form.

4.3.2 *Morphemic*

For each word form its morphemic composition is shown as: prefix+stem+infix+suffix.

4.3.3 *Hyphenation*[4]

We hyphenate all the 1,200,000 word forms of the Morphological Lexicon in accordance to linguistic hyphenation rules (of Modern Greek grammar). The hyphenator rules are separated in two categories: in those that were handcrafted according to the rules of hyphenation in MG grammar, and in those that were produced automatically based on hyphenation information incorporated in the lexicon. The rules of the second category enable the hyphenator to cope effectively with twenty-six vowel combinations, which in some words split during syllabification and in others not. Additionally, the verbal types that are liable to produce hyphenation errors as a result of the application of the hyphenation rules, have been incorporated in a list of exceptions. This list contains about 2,700 word forms containing vowel combinations, the syllabification of which leads to sense ambiguity (Tsalidis *et al.*, 2002). For example,

ή-πια (I drank) and *ή-πι-α* (smoothly), *λό-για* (the words) and *λό-γι-α* (literary), *βιά-ζω* (to force) and *βι-ά-ζω* (to rape), *δό-λιος* (miserable) and *δό-λι-ος* (insidious).

4.3.4 *Stylistic (level of style)*

Some word forms are attributed a special usage (style) other than their default normal usage. The stylistic attributes encoded in the lexicon are: spoken, slang, archaic, dialectical, formal, and informal.

4.3.5 *Terminological (thematic domains)*

In the lemma level there is additional information denoting if the lemma belongs to any of the special thematic vocabularies (see Fig. 3).

4.3.6 *Declination*

Declination is determined by an ordered set of suffixes (endings) attached to the stem in order to produce the correct word forms. For instance, for verbs, the usual endings are {*-ω*, *-εις*, *-ει*} e.g. [*δίν-ω*, *δίν-εις*, *δίν-ει* = I give, you give, he/she, it is gives, etc.] and for adjectives, some usual endings are {-os, *-η*, *-ό*} e.g. [*καλ-ός*, *καλ-ή*, *καλ-ό* = good (in three genders)]. The electronic lexicon makes use of 191 suffix rules.

4.3.7 *Accent*

Each word with more than one syllable carries an accent mark. Its position is determined by nineteen accent rules encoded in the lexicon, thus producing the correct accented orthographic transcription of every word form. Accent position is lexical, that is, it may vary during declination and it contributes to lexical identity. Many word pairs exist that differ only in accent, for example: [*γέρος* = old man], [*γερός* = strong].

4.3.8 *Alternative, parallel, or derived word forms*

In this information level the lexicon specifies related/derived word forms. Using the tags: ancient word (ANCIENT), foreign word (FOREIGN), informal word (INFORMAL), related words (REWORD) we link different lemmas.

4.4 LexEdit: the Lexicon Editor

In order to automate and simplify the coding of lexical information, a special tool (LexEdit) (Fig. 4) was developed. LexEdit is a Lexicon Editor. Figure 5 shows a typical screen capture of LexEdit displaying processed lexical entries. In the left pane of the application window we can see the sections that incorporate lemmas of the lexicon. We have a section for each Greek alphabet character. The 'iota' section is selected and in the right pane we have a part of the lemmas starting with the Greek character 'iota'. The information presented in the detailed view of the right pane is: (i) the lemma or label in the first column, (ii) the morphology of the lemmas, i.e. the constituent morphemes (except the suffix), (iii) the number of meanings, (iv) the part of speech (POS) and (v) a description (or comments) in the last column. The notation used for the morphology representation is: < > surround prefixes, {} surround stems, [] surround infixes. In Fig. 5 we can see composite words with more than one stems as well as more than one prefixes and infixes.

In LexEdit default lists are used for accents and for endings. So the editor provides two alternatives: (i) the automatic pre-selection from lists (through the inflections and their combinations) and (ii) the manual (custom rules), where the (human) lexicographer has the ability to specify his own combination for new lexical word forms that might have emerged due to the evolution of the language. However, the lexicography editor is outlined in a way (separates accent suffixes, suffixes rules) that facilitates the creation of new grammar rules.

4.5 Internal structure of the lexicon

The internal structure of the lexicon makes heavy use of tag lists (i.e. pre-selected lists of characteristics) and rules determining the attribution of the appropriate tags to every input word. This approach facilitates both debugging and expanding the lexicon, since the lists can be easily updated/enriched and reloaded into the system.

The tags and the rules are applied successively in a five-layer model (see Fig. 6) as described below.

Humanistic		*Technological*	
archaeology	literature	Logistic	business management
religious	history	architecture	biology
music	philosophy	agronomy	astrology
movies	philology	anthropology	economy
ethnology	painting	genetics	electrology
folklore	rhetoric	botany	electronics
dancing	sociology	chemistry	cosmography
pedagogy	mythology	meteorology	mathematics
theatre	politics	mechanical engineering	graphistics
theology	linguistics	computer science	physics
psychology	sculpture	medical science	technology
science of religion	printing	geography	geology
photography	pastry-making	military terms	mineralogy
journalism	cooking	ecology	zoology
athletics	cosmetics	mariner terms	

Fig. 3 Thematic Domains.

4.5.1 *First layer*

Employs a tag set of sixty-seven unique attributes determining part-of-speech and grammatical information. The tags are applied directly to the lemma but any value stored in this layer will automatically be inherited to every derived word form as well. Also in this layer an additional set of seventy-nine domain attributes and style attributes assign more specific information on the lemma relating it either to the thematic field (i.e. tag attribute: archaeology) or to the stylistic information (i.e. tag attribute: spoken) or to the usage (i.e. tag attribute: FOREIGN). The tag attributes are used to link the defined (source) lexeme with another (destination) lemma. Actually, we can link a specific meaning of the source lemma, with a specific meaning of the destination lemma. We can also specify that the link refers to a subset of the word forms from source lemma to a subset of word forms in destination lemma, defining the *from* and *to* attributes, respectively. For example, if a lemma embodied in Greek comes from another language, this lemma is linked with the equivalent Greek lemma: [κουμπιούτερ=computer] is linked with the Greek lemma [υπολογιστής=computer] by the tag attribute (FOREIGN).

The morphological processor contributes some additional tags in this layer that specify information about foreign words, synonyms, homophones, or lemmas leading to semantic ambiguity.

4.5.2 *Second layer*

On this layer, the lexicographic introduces seventy-seven morphological characteristics (attribute tags), e.g. [ACC_PLUR] for accusative and plural, or [B_P_SPOKEN] for second person and

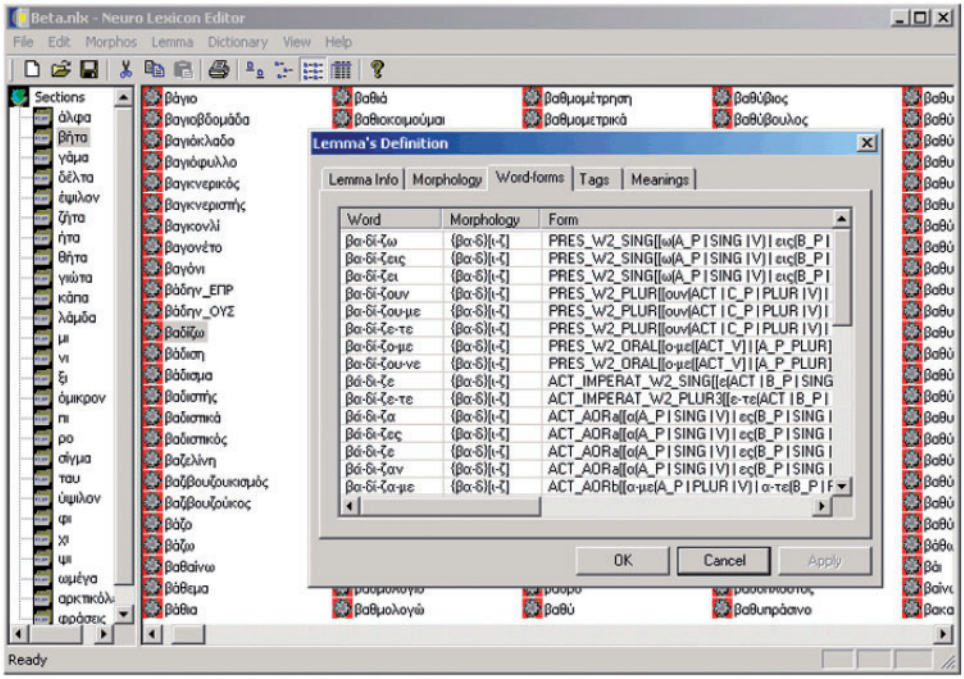

Fig. 4 Snapshot of the lexicographic editor.

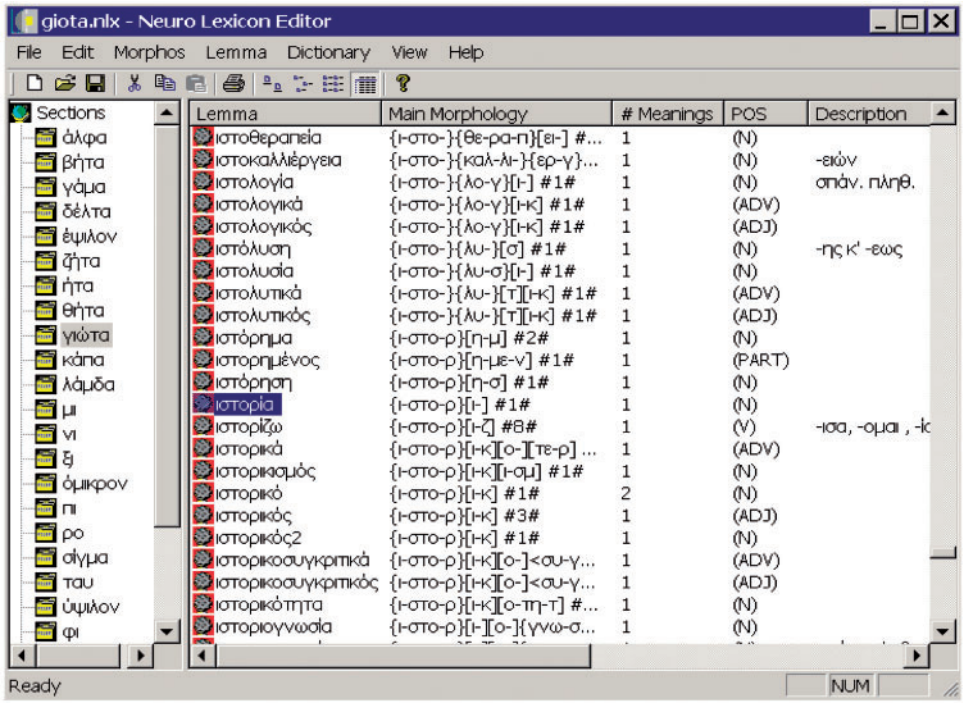

Fig. 5 The Neurolingo Lexicographic Editor (LexEdit).

colloquial Greek. For each attribute tag we assign analytically the morphological characteristics. For example in attribute tag [ACC_PLUR] we encompass the unique attributes [ACC=accusative] and [PLUR=plural]. These morphological characteristics will later describe wider groups of words (all the words that have the morphological characteristics accusative and plural).

4.5.3 *Third layer*

This layer specifies the location of accent using nineteen accent rules. The accent rules specify both the initial accent location and the accent movement during the generation of inflected word forms. So (i) for word that is written without accent like [γη = earth], we have the accent information {NO_ACCENT}, (ii) for the group of words where accent moves from the antepenultimate syllable to penultimate and then to antepenultimate again (e.g. [άν-θρωπος, ανθρώπου, άν-θρωπο]), we have the accent rule {UNTER_PEN = antepenult, penult, antepenult}.

4.5.4 *Fourth layer*

Here, we specify the suffixes using 191 suffix rules. For example the suffix {-α} is the suffix with morphological attribute values: first person, singular, verb or nominative, plural for neuters. Wider suffix rules (grouping many suffixes) are also specified, in order to facilitate wider inflected generation. For example, for perfective aspect of consonant ending verbs with syllabic augment we specify the rule {ACT_AORa} in which we describe the first singular person's suffix with ending {-α} e.g. [έπαιξ-α = I played], the second singular person's suffix with ending {-ες} e.g. [έπαιξ-ες = you played], the third singular person's suffix with ending {-ε} e.g. [έπαιξ-ε = he/she, it played] and the third plural person's suffix with ending {-αν} e.g. [έπαιξ-αν = they played], with the related morphological characteristics (attribute names). This specification enables better search for lemma stems. Below we give short examples of suffixes for verbs and nouns:

e.g. Lexeme [παί-ζω=I play] <παι-ζ>
\suffix_rule{W}, \suffix{ω} {\attribute_name
{A_PERSON, SING, PRESENT TENSE, FUTURE TENSE}{2}
\suffix_rule{A}, \suffix(α) {\attribute_name
{A_PERSON, SING, PERFECTIVE ASPECT, HABITUAL ASPECT{2}
e.g. Lexeme [πό-λη=city] <πο-λ>
\suffix_rule{H}, \suffix {η} {\attribute_name
{SING, FEM, NOM, ACC, VOC}{2}
\suffix_rule{HS}\suffix {ης} {\attribute_name
{SING, FEM, GEN}{2}

For each lemma we generate all the word forms that are still in use, including the words with different suffixes according to style information (dialectic word form, archaic word form); for example the words [πό-λε-ως[5] = city's], [α-πο-φά-σε-ως = decision's], [α-να-λύ-σε-ως = analysis's] etc., have the suffix -εως that is still in use in Modern Greek in many female nouns of this category (Triantafyllidis, 1991). This suffix (-εως) is not generally used but is used only in definite lemmas having this suffix in oral and written speech.

We do not generate the periphrastic tenses as a set, whereas we generate their components: i.e. the auxiliary verb [έχω=have], [είχα=had], or the particle [θα=will] as well as the infinitive.

4.5.5 *Fifth layer*

Given that we have: (i) the morphological information (attributes), (ii) the accent position and (iii) the suffix values, on the last layer we designate the 306 grammar rules for the generation of group categories of lemmas, like those designated by traditional grammar (i.e. masculine antepenultimate noun with ending -ος, or present tense of penultimate verb). To give an example, for the generation of nouns with feminine gender and ending (-η), the description has the following form:

\grammar_rule{ADJ_FEM_H1}
{\attribute_name{ADJ}{0}}
{\suffix_value{FEM_H_SING} {}
\accent_value{ULT1} {}
\attributes
{\attribute_name{INFORMAL}{0}}}
{\suffix_value{PLUR_ESWN}
{}\accent_value{ULT1} {}
\attributes
{\attribute_name{FEM}{0}}}}

Schematically the architecture is sketched in Fig. 6. If we have the description of all grammatical

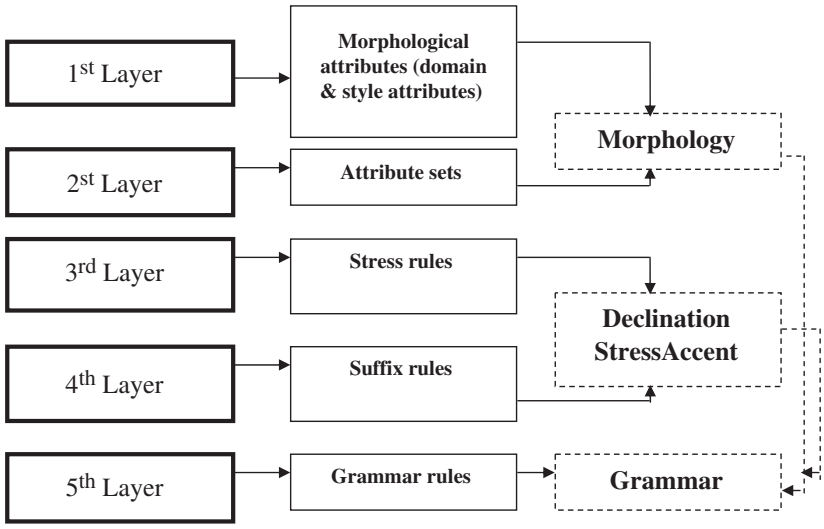

Fig. 6 Table of levels of morphological lexicon.

rules in lexicographic editor (Lex Edit), the description of lemmas is (in screen captures) as shown in Figs 7–9.

Lexeme [μα-γα-ζί=shop]
{μα-γα-ζ} {{N_NEUT_I1{NEUT_I_SING [[ι(ACC | NEUT | NOM | SING | VOC) | ιου(GEN | NEUT | SING)] ()] ULT1<<Λ(1)()>()>() || NEUT_I_PLUR[[ια(ACC | NEUT | NOM | PLUR | VOC) | ιων(GEN | NEUT | PLUR)] ()]ULT1<<Λ(1)()>()>()}(N)}}.
{ma-ga-z}[a-k] {{N_NEUT_I2_NOGEN {NEUT_IIA[[ι(ACC | NEUT | NOM | SING | VOC) | ια(ACC | NEUT | NOM | PLUR | VOC)] ()]PEN1<<Π(1)()>()>()} (N)}})Moreover, the information that {μα-γα-ζ} is the stem and the [α-κ] is infix is given. That is to say that each lemma contains all stems that belong grammatically to the same category.

The word production has the output as shown in Fig. 10.

5 Data Structure of the Neurolingo Lexicon

There are two variations of Finite State Automata [FSA] (Hopcroft and Ullman, 1979), that have been thoroughly used as lexical representation structures in the Neurolingo MRD: (i) the Minimal Directed Acyclic Graphs [MDAGs] (Hopcroft and Ullman, 1979) (Lucchesi and Kowaltowski, 1993) and (ii) the TRIEs (Knuth, 1973), (Aho *et al.*, 1983). Their capacity to store and manipulate large word sets is complemented by several additional features concerning:

- Speed. The speed of the lookup function depends on the length of the searched word and not on the size of the lexicon.
- Sorting Convenience. The words stored in an FSA can be easily sorted, by sorting the outgoing transitions of each node.
- Regular Expression Support. An FSA can easily evaluate complex regular expressions. This also permits the development of smart word correction algorithms, which utilize regular expressions to produce alternative words.

Both MDAG and TRIE represent common prefix paths. MDAG also represents common suffix paths, resulting to smaller automata (fewer states and transitions). We are using MDAG to store the words of our spelling lexicon and TRIE to index the lemmas of the morphological lexicon. More than 1,000,000 Modern Greek word forms (a 12 MB text file size) were converted to an MDAC structure

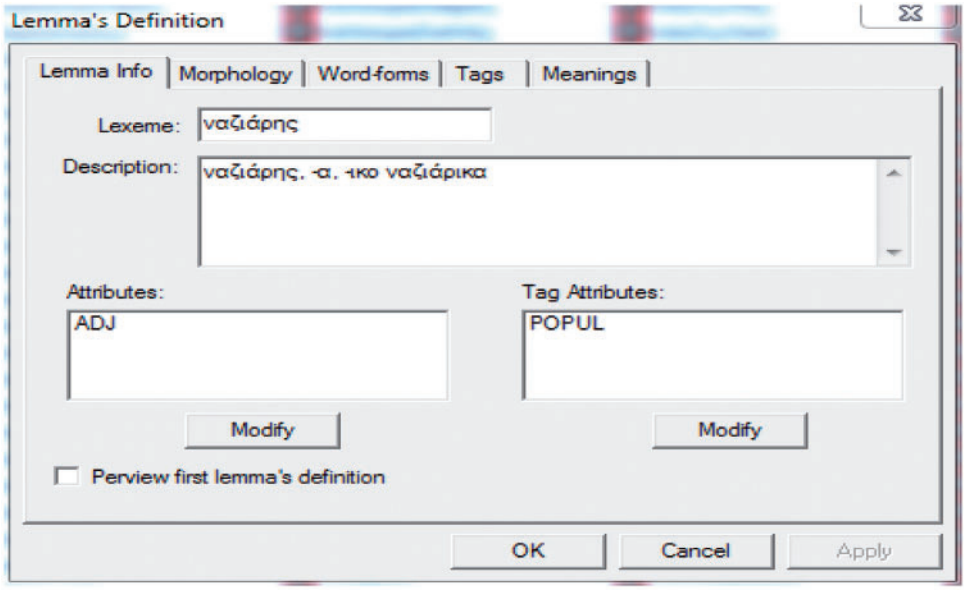

Fig. 7 Domain and style attributes (1st and 2nd level).

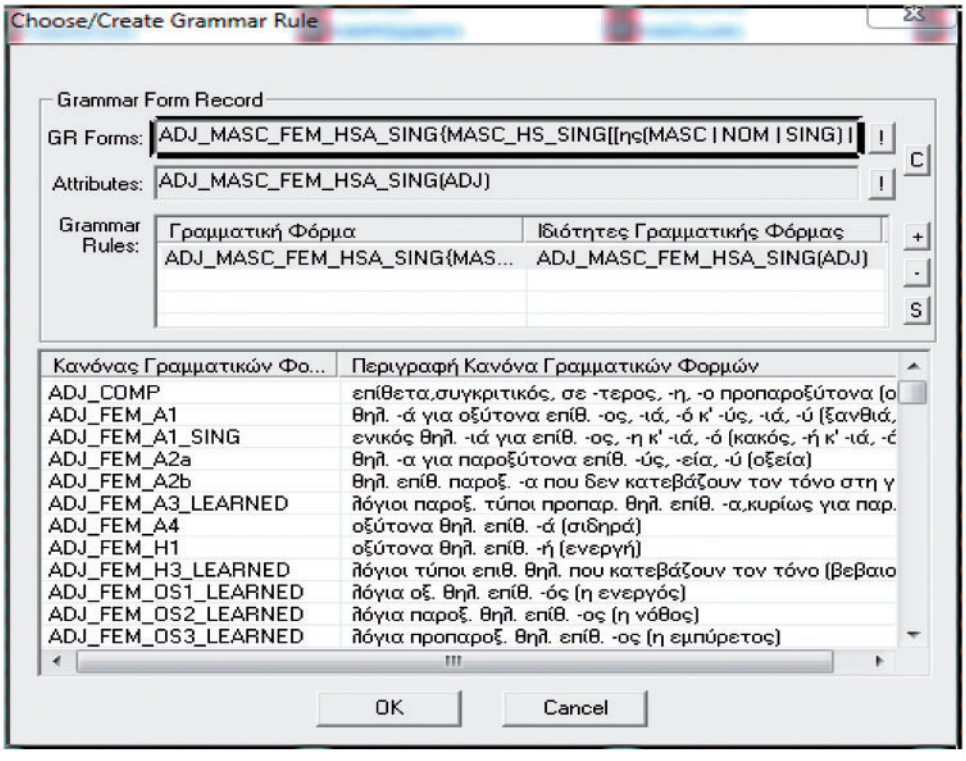

Fig. 8 Domain of accent and suffixes rules (3rd layer).

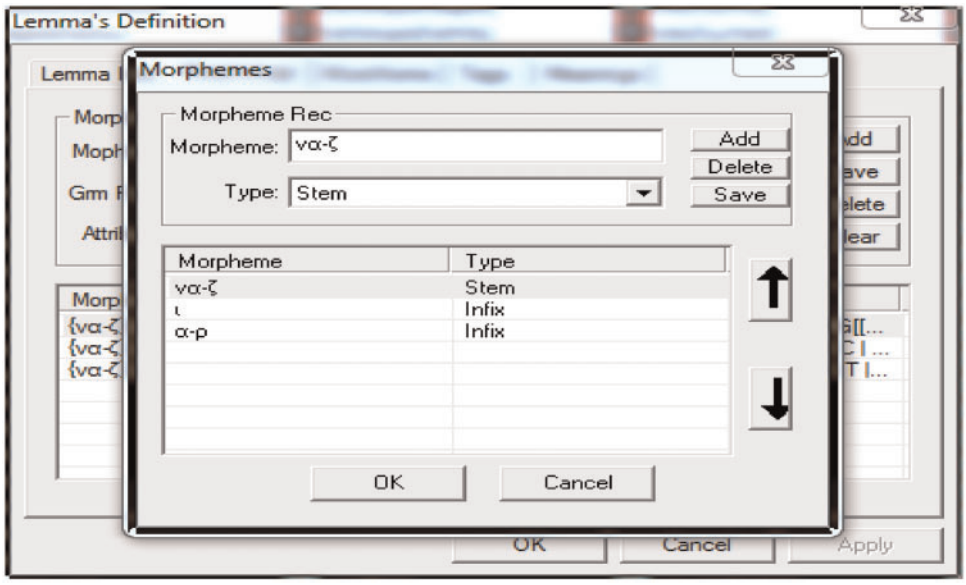

Fig. 9 Morpheme information.

Word forms	Morphology	Form	Attribute
μα-γα-ζί ma-ga-zi	{μα-γα-ζ} {ma-ga-z}	ι(ACC | NEUT | NOM | SING | VOC..	N_NEUT
μα-γα-ζιού ma-ga-zioy	{μα-γα-ζ} {ma-ga-z}	ιου(GEN | NEUT | SING)…	N_NEUT
μα-γα-ζιά ma-ga-zia	{μα-γα-ζ} {{ma-ga-z}	ια(ACC | NEUT | NOM | PLUR | VOC..	N_NEUT
μα-γα-ζιών ma-ga-zion	{μα-γα-ζ} {ma-ga-z}	ιων(GEN | NEUT | PLUR..	N_NEUT
μα-γα-ζά-κι ma-ga-za-ki	{μα-γα-ζ}[α-κ] {ma-ga-z}[a-k]	[[ι(ACC | NEUT | NOM | SING | VOC..	N_NEUT DIM
μα-γα-ζά-κια ma-ga-za-kia	{μα-γα-ζ}[α-κ] {ma-ga-z}[a-k]	| ια(ACC | NEUT | NOM | PLUR | VOC)..	N_NEUT DIM

Fig. 10 Word production.

(of size 790 KB), following a method similar to that of Mihov (Mihov, 1998). The search speed of this structure is approximately 600,000 words/second on a 1.7GHz MS Windows computer. For 26,000 lemmas, using approximately 400,000 word forms as indexing keys, the size of the TRIE index is 5.6 MB.

6 Applications and Uses of the Neurolingo Lexicon

The presented lexicon was designed specifically for Modern Greek and contains complete morphological and grammatical information for all POS of the language. The lexicon is a language resource utilized by all Neurolingo's language tools. Specifically: (i) Neurolingo Hyphenator's ability to handle the phenomenon of synizesis is based on knowledge extracted from the lexicon's syllabification information, (ii) Neurolingo Speller's functionality is based on the lexicon's orthographic information, (iii) Neurolingo Lemmatizer's functionality is based on an index that contains all the lexicon's word forms, in order to normalize any word form to the corresponding lexical unit, (iv) Neurolingo Thesaurus Browser's ability to handle the morphological variation of search terms in user's queries is based on an index that contains all word forms of each Thesaurus lemma. Moreover, Thesaurus utilizes the morphosyntactic information contained in the electronic lexicon in order to return synonym/antonym word forms that have the same morphosyntactic attributes as the search term.

The Neurolingo proofing tools suite (including Hyphenator, Speller, and Thesaurus) based on this Morphological lexicon constitute a functional component of almost all contemporary office suites and text processors (i.e. Microsoft Office, Open Office Writer, Lotus Word Pro) and professional publishing systems (i.e. Adobe InDesign, Quark Xpress). As for the Lemmatizer, it is used as a plug-in for Microsoft SQL Server and MySQL server.

The presented lexicon has also been used as a linguistic resource in several research projects (IATROLEXI[6], Meta-on[7], education of Muslim children,[8] etc.) (Tsalidis *et al.*, 2003; 2004a; 2004b).

7 Conclusion

In this article, we have introduced the Neurolingo electronic lexicon for Modern Greek. The Neurolingo (MRD) lexicon incorporates a morphological processor and is able to recognize and generate approximately 1,200,000 word forms from its approximately 90,000 stored lemmas. The tool (processor plus lexicon) is available both as a platform-independent java library (.jar format) and in native form (C++ API for Windows, Unix, Mac OS X).

It can serve as a functional component of almost all contemporary text processors (e.g. MS Word, Open Office Writer, Lotus Word Pro) and professional publishing systems (e.g. Adobe InDesign, Quark Xpress). It helps users (typists, typesetters, writers, translators, editors, etc.) to carry out automatically most time-consuming text processing operations. This tool is used in the following Neurolingo products: (i) Proofing Tools (Speller, Thesaurus, Hyphenator) for Windows MS Office[9], (ii) Proofing Tools for OpenOffice and StarOffice,[10] (iii) Proofing Tools for MS Office for Mac OS,[11] (iv) Speller and Hyphenator for Quark Xpress,[12] (v) Speller and Hyphenator in the Adobe products[13] (vi) Lemmatizer for MS Windows (SQL Server and Index Services) and other index software (Apache lucene).[14] The Speller, the Hyphenator and the Thesaurus for MS Office 2000/XP/2003 (Windows) and X/2004 (Mac OS X) are bundled together in one CD-ROM and the buyers of the product have free access to the following services: (a) Download updates of Proofing Tools for MS Office (when available), (b) Download and try the beta version of the Speller for Polytonic Modern Greek for MS Office, (c) Access the Hyphenator and the Thesaurus from within MS Office Research Services, (d) Download the Proofing Tools for NeoOffice 2.0.3 for Mac OS X on Intel platform and (e) Download other new versions of Proofing Tools for Open/Star/Neo Office (when available).

Moreover, through the online free compound language tool (Lexiscope[15]) the user can check the functionality of Neurolingo's Lexicon, Hyphenator, Speller, Lemmatizer, and Thesaurus.

In the future plans of Neurolongo, the presented MRD is going to support the Modern Greek Grammar Checker (another NLP tool currently under development by Neurolingo), which will be able to detect syntactic disagreement, semantic ambiguity, declination, and syntactic errors, and generally errors that demand more elaborate management than a simple speller.

References

Aho, A.V., Hopcroft, J.E., and Ullman, J.D. (1983). *Data Structures and Algorithms*. Reading, MA: Addison-Wesley, pp. 163–9.

Babiniotis, G. (1998). *The Dictionary for the Modern Greek*. Athens (in Greek).

Baltzis, S., Kolalas, S., and Eumeridou, E. (2005a). Computational modern Greek morphological LEXICON – An efficient and comprehensive system for morphological analysis and synthesis. *Literary and Linguistic Computing*, **20**(2): 153–87.

Baltzis, S., Kolalas, S., and Eumeridou, E. (2005b). *Computational Morphological Lexicons – Current tools for Knowledge and Communication Management, Proceedings of the 2nd Balkan Conference in Informatics*. Ohrid – FYROM, pp. 346–56.

Boguraev, B. and Pustejovsky, J. (1990). *The Role of Knowledge Representation in Lexicon Design, Proceedings of the 13th conference on computation linguistics*, pp. 36–41, inland.

Carr, M. (1997). Internet Dictionaries and Lexicography. *International Journal of Lexicography*, **10**(3).

De Schryver, G.-M. (2003). Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography*, **16**(2): 143–99.

Detorakis, Z. (2009). *Development of Algorithms for Natural Language Processing*. Ph.D. thesis, National Technical University, Athens.

Fytrakis, T. (1997). *The Major Greek Dictionary*. Athens (in Greek).

Hanks, P. (2001). *The Probable and the Possible: Lexicography in the Age of the Internet, Asialex Proceedings*, pp. 7–36.

Hopcroft, J.E. and Ullman, J.D. (1979). *Introduction to Automata Theory, Languages and Computation*. New York: Addison-Wesley.

Iordanidou, A. (1991). *Verbs of Modern Greek*. Athens (in Greek): Patakis Edition.

Knuth, D.E. (1973). The Art of Computer Programming. *Sorting and Searching*, vol. 3. Reading, Massachusetts: Addison-Wesley, pp. 481–505.

Kriaras, E. (1995). *Greek Dictionary of the Modern Demotic Language*. Athens (in Greek): Ekdotiki Athinon Editions.

Koskenniemi, K. (1983). Two-level morphology: A general computational model for word-form recognition and production. *Department of General Linguistics*. Publication No.11, University of Helsinki.

Lucchesi, C. and Kowaltowski, T. (1993). Applications of finite automata representing large vocabularies. *Software Practice and Experience*, **23**(1): 15–30.

Lyons, J. (1977). *Semantics*, **vol. 1 & 2.** Cambridge University Press.

Mihov, S. (1998). Direct building of minimal automaton for given list. *Faculté de Mathématiques et Informatique*, **91,** Sofia, Bulgaria, Livre 1 Editions.

Ntais, G. (2006). *Development of a Stemmer for the Greek Language*. Master Thesis, Department of Computer and Systems Sciences, KTH-Stockholm University.

Papageorgiou, H., Prokopidis, P., Giouli, V., and Piperidis, P. (2000). A Unified POS Tagging Architecture and its Application to Greek. *RLEC*. Athens, pp. 1455–63.

Perry, B.C. (1997). Electronic learners' dictionaries (ELDs): an overview of recent developments. *CALL Electronic Journal*, **1**, http://www.jaltcall.org/cjo/5_98/call_EJ/Perry.html.

Petasis, G., Karkaletsis, V., Farmakiotou, D., Samaritakis, G., Spyropoulos, C., and Androutsopoulos, I. (2001). *A Greek Morphological Lexicon and its Exploitation by a Greek Controlled Language Checker, 8th Panhellenic Conference on Informatics*, **vol. 1**. Cyprus, pp. 80–89.

Petasis, G., Karkaletsis, V., Farmakiotou, D., Androutsopoulos, L., and Spyropoulos, C.D. (2003). *A Greek Morphological Lexicon and Its Exploitation by Natural Language Processing Applications, Advances in Informatics - Post-proceedings of the 8th Panhellenic Conference in Informatics*, pp. 401–19.

Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D., and Androutsopoulos, I. (2000). Using machine learning techniques for part-of-speech tagging in the Greek language. In Fotiadis, D.I. and

Nikolopoulos, S.D. (eds), *Advances in Informatics*. Singapore: World Scientific, pp. 273–81.

Sgarbas, K., Fakotakis, N., and Kokkinakis, G. (1995). A PC-KIMMO-Based Morphological description of Modern Greek. *Literary and Linguistic Computing*, **10**(3): 189–201.

Sgarbas, K., Fakotakis, N., and Kokkinakis, K. (1999). *Morphological Description of Modern Greek by using the two-level-model, Proceedings of the 20th Annual Meeting of the Department of Linguistics, Thessaloniki*. Greece, pp. 419–33.

Sgarbas, K., Fakotakis, N., and Kokkinakis, G. (2000a). Two algorithms for incremental construction of directed acyclic word graphs. *International Journal on Artificial Intelligence Tools*, **4**(3): 369–81.

Sgarbas, K., Fakotakis, N., and Kokkinakis, G. (2000b). *A Straightforward Approach to Morphological Analysis and Synthesis, Proceedings COMLEX 2000, Workshop on Computational Lexicography and Multimedia Dictionaries*. Greece, pp. 31–4.

Tambouratzis, G. and Carayannis, G. (2001). Automatic Corpora – based stemming in Greek. *Literary and Linguistic Computing*, **16**(4): 445–66.

Triantafyllidis, M. (1998). Institute of Greek studies of Aristotle University of Thessalonica. *Dictionary of common Modern Greek language*. Thessaloniki: Manolis Triantafyllidis Foundation (in Greek).

Triantafyllidis, M. (1991) [1941]). *Modern Greek Grammar*. 3rd rev. ednThessaloniki: Manolis Triantafyllidis Foundation (in Greek).

Tsalidis, C., Orphanos, G., and Iordanidou, A. (2002). *[Did the Greek computers learn how to hyphenate(?)]. Studies in Greek Linguistics, Proceedings of the 23rd Annual Meeting of the Department of Linguistics*. Thessaloniki, Greece: Aristotle University of Thessaloniki, pp. 901–911.

Tsalidis, C., Vagelatos, A., and Orphanos, G. (2004a). *An Electronic Dictionary as a Basis for NLP tools: The Greek Case, 11th Conference on Natural Language Processing* (TALN '04) Fez, Morocco, April 19–22.

Tsalidis, C., Orphanos, G., Iordanidou, A., and Vagelatos, A. (2004b). *Proofing Tools Technology at Neurosoft S.A.* Workshop on International Proofing Tools and Language Technologies, Patras, Greece, July 1–2.

Tsalidis, Ch., Vagelatos, A., and Orphanos, G. (2003). *Implementation of a Greek Morphological Lexicon for the Biomedical Domain*. International Conference on Information Communication Technologies in Health (ICICTH 2003), Samos, Greece.

Tziafa, E. (2007). *Construction of an Electronic Dictionary of Financial and Stock Market Terms – General Comments*. 6th Congress, (ELETO).

Van Eijck, J. and Jaspars, J. (1996). *Ambiguity and Reasoning*. Technical Report CS-R9616. CWI, Amsterdam: Dutch National Research Institute for Mathematics and Computer Science.

Notes

1 Neurolingo company employs computer engineers and linguists specialized in natural language processing. Neurolingo was established in December 2005 by the members of Neurosoft Language Technology Team (http://www.neurosoft.gr).
2 http://www.ilsp.gr/parole_eng.html
3 http://linginfo.frl.auth.gr
4 The role of the *Hyphenator* is to indicate all hyphenation points of a word. This practice is not applicable to short paragraph lines (i.e. in newspaper columns) or when long words occur at the end of some lines, as excessive word spacing affects the aesthetics of text and part of the printable space is wasted in blank spaces.
5 The ending is [ε-ως] and the theme is [πολ]
6 www.iatrolexi.gr
7 www.metaon.gr
8 www.museduc.gr
9 http://www.neurolingo.gr/en/products/proofing_tools/mso.jsp
10 http://www.neurolingo.gr/en/products/proofing_tools/oo.jsp
11 http://www.neurolingo.gr/en/products/proofing_tools/mso08.jsp
12 http://www.neurolingo.gr/en/products/proofing_tools/qx.jsp
13 http://www.neurolingo.gr/en/technology/application_tools/speller.jsp, http://www.neurolingo.gr/en/technology/application_tools/hyphenator.jsp
14 http://www.neurolingo.gr/en/technology/application_tools/lemmatizer.jsp
15 http://www.neurolingo.gr/en/online_tools/lexiscope.htm