IMPLEMENTATION OF A GREEK MORPHOLOGICAL LEXICON FOR THE BIOMEDICAL DOMAIN

Ch. Tsalidis, G. Orphanos
Neurosoft S.A.
Kofidou 24, N. Ionia
14231 Athens, Greece
Tel: +30 210 2719943
Fax: +30 210 2717941
E-mail: tsalidis, orphan@neurosoft.gr

A. Vagelatos
R.A. Computer Technology Institute
Eptachalkou 13, Thiseio
11581 Athens, Greece
Tel: +30 210 3416220
Fax: +30 210 3416250
E-mail: vagelat@cti.gr

Abstract

This paper presents the extension of a Modern Greek morphological lexicon with medical information pertaining terminology and disease hierarchies, included but not limited to the 10th revision of the International Classification of Diseases (ICD-10). The utility of the lexicon is manifold: spell checking and correction of medical terms, normalization of morphological variations, disease indexing, browsing into disease hierarchies. The definition and manipulation of lexical entries are achieved with the aid of a specialized editor, called *LexEdit*. Some basic features of the underlying lexical representation mechanism are its openness to new kind of information, its ability to manage lexical units on morpheme basis and its ability to cope with the complex inflectional system of Modern Greek as well as the existence of marked stress.

Introduction

Some years ago a morphological lexicon for Modern Greek was constructed in order to be the basis of a spell checking/correction system [Vagelatos et al., 1994]. Later on, the demand for the incorporation of more specific terminology was posed especially by the Greek medical society. At the same time the Greek Ministry of Health and Welfare started a project for the translation of ICD-10 in Greek. A year later the Greek version of ICD-10 was a reality and it was decided to be incorporated into the general-purpose morphological lexicon.

In this paper we present the process of ICD-10 terms incorporation into the original Greek morphological lexicon. First, we present the main characteristics of the lexicon. Next, we describe the lexicon-coding scheme along with a specialized editor that was implemented for the manipulation of the lexicon entries. Then we present some technical figures regarding the actual stage of ICD-10 incorporation into the lexicon and finally come the conclusions as well as some future plans.

The Lexicon

The original lexicon was developed with the capability to include morphological as well as semantic and syntactic information in order to support various NLP applications, according to the following specifications [Stamison-Atmatzidi et al., 1994]:

- Each *lexical entry* is a cluster of all running word forms of a lexeme. Lexeme is the representative of the cluster (headword).
- For each *word form* its segmentation into syllables must be present.
- For each *word form* its segmentation into morphemes must be present.
- Stress must be handled with an easy and efficient way.
- We must be able to incorporate simple (property) information as well as compound (structured) information.
- We must be able to encode the meanings of a lexeme, as in printed lexicography.
- We must be able to define reference pointers between lexemes in order to represent *WorldNet* links (*WordNet* is a lexical database that is organized into synonyms sets. More information can be found at: http://www.cogsci.princeton.edu/~wn/).

In order to fulfill the above-mentioned specifications, a coding scheme was devised to represent all the necessary information. In the next section the main characteristics of this coding scheme are presented with references to *LexEdit* screenshots.

LexEdit: A Lexicon Editor

In order to automate and simplify the coding of lexical information, a special tool was constructed. *LexEdit* is a Lexicon Editor, which was used for the definition of the ICD-10 terms to be inserted in the lexicon.

Figure 1: LexEdit Application

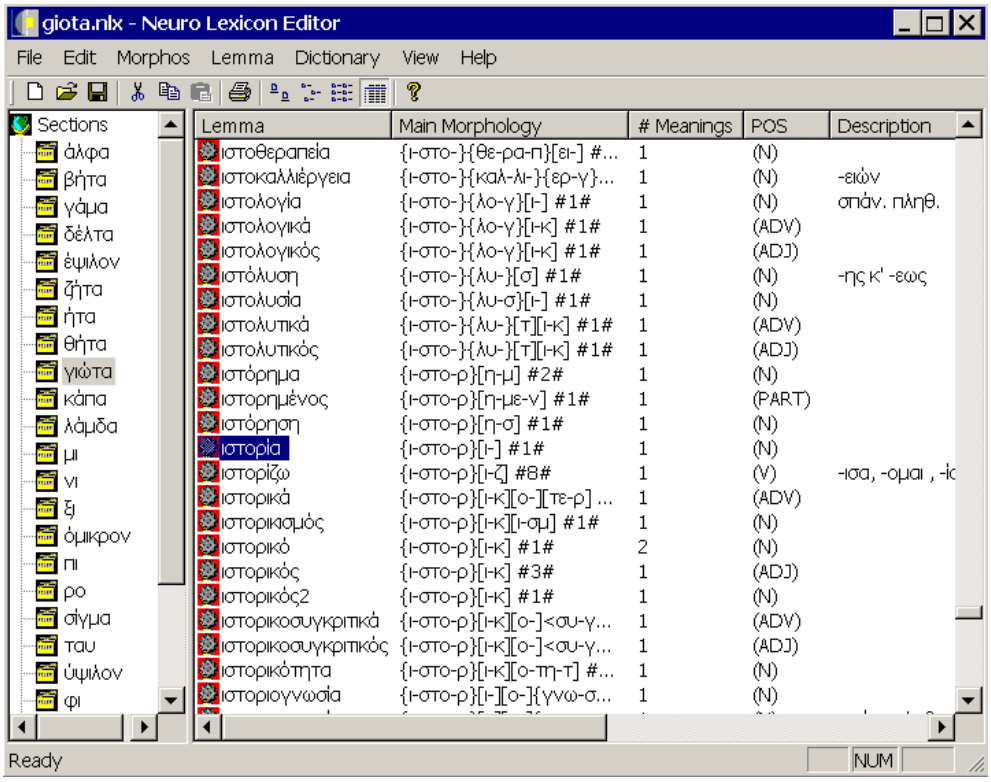


Figure 1, shows a typical screen of *LexEdit* showing processed lexical entries. In the left pane of the application window we can see the sections that incorporate lemmata of the lexicon. We have a section for each Greek alphabet character. The "iota" (γιώτα – ι) section is selected and in the right pane we have a part of the lemmata starting with the Greek character "iota".
The information presented in the detailed view of the right pane is: a) the lexeme or label in the first column, b) the morphology of the lexeme, i.e. the constituent morphemes (except the suffix), c) the number of meanings, d) the part of speech (POS) and e) a description (or comments) in the last column.
The notation used for the morphology representation is: < > surround prefixes, { } surround stems, [ ] surround infixes. In Figure 1 we can see composite words with more than one stems as well as more than one prefixes and infixes.

Lexical Items

A lexical item (**lexitem**) in *LexEdit* is a lemma definition and consists of the following parts:

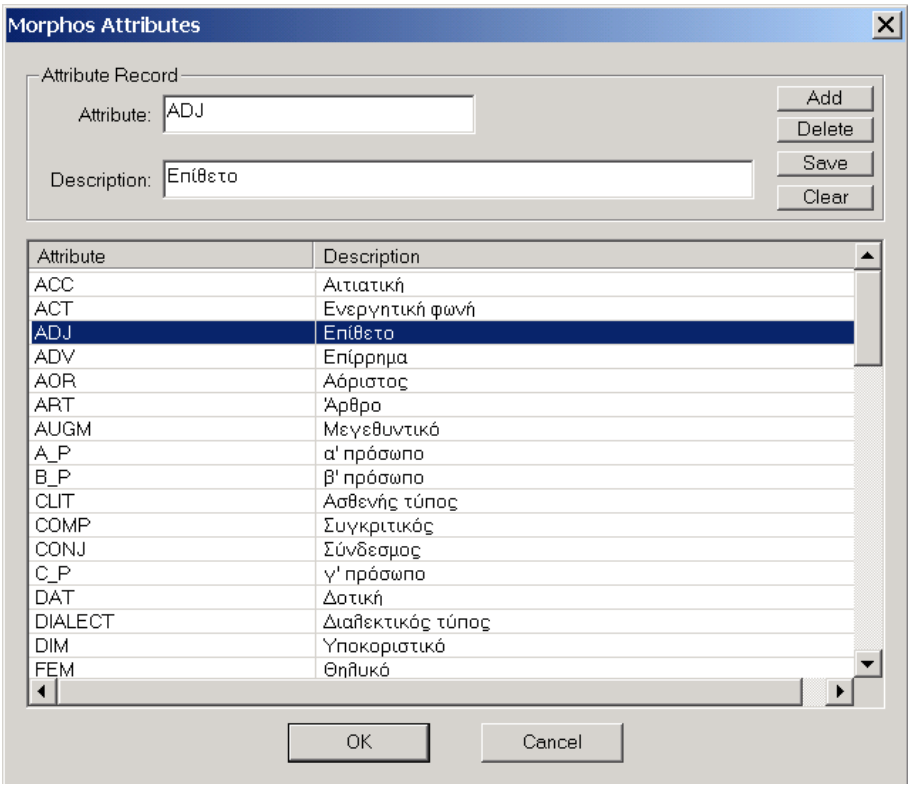
lexitem := LEX_ITEM_NAME,
descriptive-part,
informative-part, **(1)**

As formula **(1)** shows, a lexitem definition contains three parts. The name (lexeme) which characterizes the lexitem, the descriptive-part which describes how the lexeme's word forms are constructed from its constituent parts and the informative-part which holds the information that can accompany a lexeme.
The basic unit of word forms are the letters of the Greek alphabet. Despite this, words are usually divided in parts which contain information and usually characterize and give special meaning to the word. These parts are called morphemes and constitute the basic unit of word forms. We distinguish four types of morphemes: prefix, stem, infix and inflection (suffix).

Attributes

Attributes are the primitive information that can be attached to every constituent part of a lexeme. The information it carries is relative to the existence or not of the attribute and gives the user the ability to assign properties in lexemes.
We distinguish two kinds of attributes, Morpho and Tag (or Meaning). The Morpho Attributes are attributes that can accompany every morpho while Tag Attributes are attributes that are referred to the whole lexeme or one of the lexeme's meanings. In Figures 2 we can see the way we can define and maintain the Morpho Attributes space. Morpho attributes used mainly to incorporate morphological information while Tag Attributes are used to include semantic, stylistic or terminological information.

Figure 2. Morpho attributes.



Tags

Tags represent structured information with fields, parameters and intra-lexeme references. The user must define the *Tag Space* as the names of tags that a lexeme can have. Tag space is defined only for consistency checking in order to forbid duplicate tag structures due to name misspellings.

Examples of tags can be lists of synonyms, related words, related phrases, etc.

Stress, Suffix and Grammar Rules

The complexity of Greek inflectional morphology is handled with the definition of Stress, Suffix and Grammar rules.

Lemma Definition

Using the Morpho spaces we can define Greek lexemes that constitute the Lexicon's lemmata.

As already shown, from the descriptive point of view, a lexeme can have one or more morphologies, while from the informative point of view the lexeme can have one or more meanings.

These two kinds of information are defined for each lexeme using the Lemma's Definition Dialog.

Morphology

As already stated a lexeme can have more than one morphologies. As morphology we consider a string of morphemes of type prefix, stem, or infix (but not suffix). For example all Greek verbs have more that one

morphology. Morphologies are considered as the invariant part of the lexeme. This is the reason we do not include the suffix types in the morpheme's type. The suffixes constitute the variant part of the lexeme and are handled separately. The combination of a morphology with a set of suffixes and a set of stress positions produce the word forms of a lexeme. This coding permits the efficient and compact definition of lexemes using the same suffix-stress rules.

Meanings

A lexeme can have one or more meanings. The first meaning, called *Zero Meaning,* is the default meaning of the lexeme. The information for *Zero Meaning* defined separately from information of the other meanings with the first (Lemma) and third (Tags) tab. The other possible meanings of the lexeme are defined in the last (Meanings) tab. The meaning name of the *Zero Meaning* is the lexeme's name (label).

Lexeme reference used to link the defined (source) lexeme with another (destination) lexeme. Actually, lexeme reference can link a specific meaning of the source lexeme, (the meaning in which the lexeme reference appears) with a specific meaning of the destination lexeme. Furthermore we can also specify that the link refers to a subset of the word forms from source lexeme to a subset of word forms in destination lexeme, defining the *from* and *to* attributes respectively.

The Greek version of ICD-10

The "Tenth Revision of the International Classification of Diseases and Related Health Problems" is the latest in a series that was formalized in 1993 as the Bertillon Classification or International List of Causes of Death (http://www.who.int/whosis/icd10). The World Health Organization (WHO) published the systematic part of ICD-10 in 1992 and the alphabetic part in 1994.

WHO has proceeded in publications of ICD-10 in the following languages: English, French, Arabic, Chinese, Russian and Spanish. At the same time a number of publications were prepared (or are currently in preparation) in many different language.

The Greek Ministry of Health and Welfare, assigned the translation of ICD-10 to a local specialized vendor. The project started in the fall of 1998 and the first version of the Greek ICD-10 was a reality the summer of 1999 [Vagelatos A., 2001].

The next phase of the project was the auditing process: the classification was send to all the Greek Medical Societies with the request to respond with possible comments within 3 months period.

Indeed in the next three months most the Greek Medical societies reply with certain comments and recommendations. All these replies were jugged by the scientific team of the project and were incorporated into the Greek version of ICD-10, thus producing the version 2.0 (it can be downloaded from: http://www.ypyp.gr/GR/healthgr/codes/codes.htm).

Technical Realization

The original morphological lexicon contained almost 73 thousands lexemes with full morphosyntactic information.

The processing of ICD-10 files resulted in a number of almost 18,000 words. The use of the original lexicon with a comparison tool gave a number of about 5,000 lexemes that were not included in it. Then the process of the incorporation started.

The project lasted about three months. Two months took the lemmas definition and an additional month was required for checking the correctness of the new lexicon.

The five thousands lexemes that were added within the means of the current project from ICD-10, were assigned a special attribute “ICD-10” for identification purposes as well as for future needs. The next step of this project is the actual production of the spelling checker as the first linguistic tool that would be based on the new biomedical electronic lexicon.

Conclusion

LexEdit was used by a team of linguists for the incorporation of ICD-10 terms. It supported all the phases in Lexicon definition (lexemes selection => morphology => meanings). Additionally the tool incorporates the following functionality:

- Compaction and compression of the lexicon (or part of the lexicon) in order to be used from linguistic tools as spellers, hyphenators, etc.
- Check of the correctness and integrity of references.
- Import and export the Lexicon in a proprietary text format.
- Export of the Lexicon in XML format in order to be printed.

Our future plans include the following:

a) Incorporation of more biomedical terms extracted from other resources (codifications, medical corpus or other).

b) Use the morphology of the lexemes in order to make classifications of the terms based on common morphological characteristics (i.e. root, affixes, part of speech, etc).

c) Use of the lexicon as the basis for special search engines and indexing mechanisms in various data mining applications.

References

Stamison-Atmatzidi M, Triantopoulou T, Vagelatos A. and Christodoulakis D. (1994). The Utilization of An Electronic Morphology Dictionary and a Spelling Correction System for the Teaching of Modern Greek. *Computer Assisted Language Learning Journal*. Volume 7(1) 1, 185-190.

Vagelatos A. (2001). Standardization in Medical Informatics: A requierement for the Introduction of Information Systems. *Archives of Hellenic Medicine.* Volume 6, 2001. (In Greek).

Vagelatos A, Triantopoulou T, Tsalidis C. and Christodoulakis D. (1994). A Spelling Correction System for Modern Greek. *International Journal of Artificial Intelligence & Tools*. Volume 3(4), 429-450.