Use Of A Morphosyntactic Lexicon As The Basis For The Implementation Of The Greek Wordnet

A. Ntoulas[1], S. Stamou[1], I. Tsakou[2], Ch. Tsalidis[1 2], M. Tzagarakis[1 2]
and A. Vagelatos[1 2]

[1] Computer Engineering & Informatics Department, University of Patras, Greece
[2] Computer Technology Institute, Kolokotroni 3,
GR-26221 Patras, Greece

Abstract. Greek WordNet is a project aiming at developing a database of wordnets for the Greek language, structured along the same lines as the EuroWordNet project. This contribution presents the morphosyntactic lexicon, which will be used as the basis for the development of the whole project. This lexicon was developed within the framework of a spelling correction system. Later on, it was enhanced by adding syntactic information for each lemma and by using a relational database for the storage and management of the data.

1 Introduction

Last year, the idea of building the Greek WordNet succeeded in receiving a high grade in a GSRT[3] competition and as a consequence, in being financed by the Greek Government.

The partners for the implementation of the above idea were chosen carefully:

1. The *Computer Technology Institute*, with an active team on computational linguistics and a number of tools and resources already developed.
2. The *Linguistics Department of Patras University*, with a team of experts in linguistics who have been conducting extensive research on the Greek language, for the past years.
3. The *Informatics Department of Athens University*, with its computational linguistics team.
4. The *Computer Science Institute of the Foundation of Research and Technology Hellas*, with its experienced team on user interfaces and web based applications.

[3] GSRT = General Secretariat for Research and Technology is one of the Secretariats of the Ministry of Development and is responsible for the following activities in Greece: planning and execution of national policy on research and technology through the design and execution of relevant programmes, creation and activation of research and technological infrastructure, technological development, importation and exportation of technology, research and technology orientation, investigations of the consequences of research and technology on the economic, social and cultural development of the country.

5. *Pattakis Publications*, having under development an exhaustive lexicon of the Greek language.

The above partners agreed to cooperate for the implementation of the Greek WordNet, following the same lines as the *EuroWordNet* project[4]. The role of each one is well defined: *CTI* together with the *Informatics Department of Athens University* will develop the appropriate computer infrastructure for the project, as well as provide the linguistic resources. The *Linguistics Department of Patras University* will have as its primary goal to provide the necessary expertise in semantics for building the Greek WordNet. *CSI* will develop the appropriate user interfaces as well as the Web page that will host the final outcome. Finally, *Patakis Publications* will complete its lexicon, which is presently under development, so as to be used as a linguistic resource by the linguists.

In this paper, we present the morphosyntactic lexicon that is going to be used as a linguistic resource for the implementation of the above-mentioned project. More specifically, firstly, we give a small description of the EuroWordNet project since the Greek WordNet is going to be developed following the same structure. We, then, describe the Greek Morphosyntactic lexicon developed by CTI within the framework of a spelling correction system. Later on, this lexicon was enhanced by adding syntactic information for each lemma. Next, we present the design challenges of the information infrastructure that has to be implemented. Finally, we mention our future plans and some conclusions.

2 Goals of the Project

The main goal of the Greek WordNet project is the development of a large-scale lexical resource for the Greek language containing semantic relations between words organised around the notion of synsets (one or more word senses which are considered to be similar in meaning). The Greek WordNet will be built from existing resources and then stored in a database following the methodology of the EuroWordNet project, so as to be compatible with the specific project. The aim of the Greek WordNet is to represent the Greek lexicalization patterns and language-dependent differences of the Greek language, whilst using the language-independent ontology of the EuroWordNet for the classification of the Greek major concepts and words. This will enable the merging of the Greek WordNet in the EuroWordNet multilingual database, so as to strengthen the position of Greek in Language Technology and in particular in the field of Multilingual Information Retrieval.

During the realisation of the project, the participants will gain, from their participation to the project, significant experience and knowledge in the field of Language Engineering. The Greek WordNet project will also give them the possibility to develop important technical infrastructure that can be used as the basis for building further linguistic tools and techniques.

[4] http://www.hum.uva.nl/-ewn.

3 The EuroWordNet project

The EuroWordNet project is a multilingual database comprising WordNets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). Each WordNet represents a unique language-internal system of lexicalizations. The individual WordNets are organised around a set of Base Concepts, extracted by each country from already existing linguistic resources such as monolingual and bilingual lexicons, corpora, etc. Then, all the sets of Base Concepts are compared and a set of Common Base Concepts (concepts selected by at least two countries) is extracted. These Common Base Concepts, plus a selection of other meanings that are important in each language, function as the starting point for the development of the local WordNets, which are built taking into account the lexicalization patterns that are relevant to the specific languages. The structure of the individual WordNets is based on the formation of synsets, ie. sets of word meanings between which basic semantic relations, such as synonymy, hyponymy, meronymy, cause, etc. are expressed. In other words, Base Concepts function as "anchors" to which other concepts are attached.

All the data is then converted into the EuroWordNet import format and loaded to the EuroWordNet database. All WordNets all linked to an Inter-Lingual-Index (an unstructured list of concepts) interconnecting the languages so as to make possible to connect word meanings in one language to similar word meanings in the other participant-languages. Furthermore, all the languages that form part of the EuroWordNet share a three-entities top ontology of 63 semantic distinctions, as well as domain ontologies, which provides a common semantic framework for the classification of word meanings, enforcing, in this way, uniformity and compatibility of the different WordNets.

Today, that the EuroWordNet project is completed, it provides a large-scale linguistic resource which is used to improve the capacities and performance of current multilingual information retrieval systems with regards to reformulation of queries, automatic indexing and other issues of text retrieval task.

4 Greek Morphosyntactic Lexicon

Within the framework of a spelling checking/correction system for Modern Greek[5] our team created a morphological lexicon. In order for this lexicon to become efficient, an in-depth morphological analysis of Modern Greek (MG) was carried out.

The main characteristics of M.G. morphology can be summarised as follows:

- A complex inflectional system. For example, for the M.G. masculine nouns ending in "-os" /os/, there are six different inflections, whereas for the present tense of the active voice of verbs ending in "$-\omega$" /o/ there are seven different inflections.

[5] The spelling correction system was developed by our team at CTI, funded by Intracom S.A. and was approved and adopted by Microsoft corp. as the Greek spelling checker that will be marketed with their products.

- The existence of marked stress: Words in M.G. are stressed in either the final, penultimate or antepenultimate position.
- A "graphematic" spelling system consisting of single graphemes, compound graphemes and grapheme equivalents.
- Free-word-order: the sentence constituents (i.e. subject, verb, object) can be found in various positions. This characteristic is mainly a consequence of the inflectional system, since the syntactic role of word-forms is usually denoted by the syntactic information attached to their inflectional endings and it is not necessary for the word forms to occupy a special position within the sentence itself.
- The existence of numerous characteristics carried over from Ancient Greek. The use of old and new word forms, which gives rise to an endless linguistic debate whether both forms should continue to be used (accepted).

M.G. word classification incorporates two basic categories: The inflected and the non-inflected. Our main task was to study the inflected M.G. words in order to describe their declinations and conjugations as economically as possible.

All inflected words were given the following morphological description:

WORD = [PREFIX] + STEM + [INFIX] + INFLECTION(S)

where: STEM and INFLECTION(S) are features that are necessary for the derivation of all possible forms, PREFIX and INFIX are not always present and thus appear bracketed "[ ]" in the above description; e.g.,

/ksana-graf-tik-a/ = PREFIX(ksana) + STEM(graf) +
INFIX(tik)+INFLECTION(a)
/e-graf-a/ =PREFIX(e) + STEM(graf) +
INFLECTION(a)

This analysis led to the development of a description-language that coded both the inflectional morphology and marked stress of M.G. This language utilises 260 rules. Given the stem of every declinable word, the appropriate set of rules is attached, making possible the production of all valid forms for the particular word. Every such combination is included in the lexicon, forming a lexicon entry. This way, all words with the same stem are stored as a single lexical item along with rules regarding the allowable inflection, thus saving vast amounts of storage.

Later on, the lexicon was enhanced by extending the previously described formalism in at least 4 points:

1. The morphemes used for the formation of Greek words can incorporate sets of morphosyntactic or semantic attributes.
2. A lemma can be defined as a cluster of word-form definitions.
3. Two-dimensional information structures can accompany a lemma.

4. Syntactic or semantic relations between lemmas can be defined as typed links that are resolved during lexicon construction.

This extended formalization allows us to denote:

- Attributes: Morphological or syntactic. These attributes can accompany a morpheme or a word; e.g. %attributes = (NOUN | ARTICLE | ...).
- Infos: The name and structure of the two-dimensional tags that a lemma can use to incorporate structural or free text information; e.g. %infos = (ANTONYM: ... | MEANING: ... | ...).
- Attribute sets: Sets of attributes that can accompany a morpheme or a word; e.g. @MCSING = (MASCULINE | SINGULAR).
- Inflection: Suffixes for gender, number, case, person, etc. Inflection is denoted by inflectional rules; e.g. #MSCas$ = (as @ (NOMINATIVE) | a @GAV) @ MSCSING "masculine singular ending in -as".
- Stress: Words in M.G. are stressed on the final, penultimate or antepenultimate position while single-syllable words are not stressed. Stress is denoted by stress-rules; e.g.
  !a1 = (1). – stress in final position
  !a2 = (2). – stress in penultimate position
  a6 = (3). – stress in antepenultimate position
- Forms: Combination of infix, suffix, stress and either attributes or attribute sets, e.g. $S1 = (o-te-r #OSe !a6 @(MASCULINE
  | – o-te-r = Infix,
  | – #Ose = infl. Rule
  | – !a6 = stress rule
  | – @(MASCULINE)=attribute
  o-te-r #OSp !a6 @(MASCULINE) |
  o-te-r #THIi !a6 |
  o-te-r #ESWN !a6 @(FEMININE) |
  o-te-r #OUDOe !a6 |
  o-te-r #OUDOp !a6 |) @EPSIGR
  – @EPSIGR = attributes applied – to all morphemes.

Our morphological lexicon consists of about 80,000 entries, whereas all possible forms produced reach about one million. The primary indexing / storage mechanism used to access the words of the lexicon is the Compressed Trie. This data structure was chosen as the most appropriate for our purpose of creating a lexicon, since it makes possible efficient searching and occupies minimal disk space. The Compressed Trie is used as an index to the database records of the lemmas, and since it is relatively small (about 700Kb), compared to the size of data needed to represent the entire lexicon, it is possible to load it in the main memory as a whole. A path in the Trie, starting from the root and ending to a leaf, represents a word prefix; the leaf points to those lemma records in the database that contain the suffixes of word-forms with this prefix. Using the Compressed Trie as an index to access the lemma records, we achieved approximately one disk access per search for data located in the disk.

5 Information Infrastructure and Linguistic Problem Domains

Apart from providing the morphosyntactic lexicon for the development of the Greek wordnet, CTI is responsible for the implementation of certain linguistic tools, thus creating the appropriate information infrastructure for the linguist to work with.

The role of such infrastructure is to provide all the necessary structures and operations that will allow the linguist to operate upon them. In order to capture the linguistic problem domain better, we outline the usage of such a system and identify the paradigms and patterns on which it is based.

Information Analysis. Linguists, especially those working in the field of lexicography, do not have a clear overview or understanding of the size of the problem. Lexicographical lemmas can be very simple (e.g. a simple link to a lemma explaining the word) or complicated structures that include synonyms, example sentences, references to other entries in the lexicon etc. In order to code such entries, linguists need tools that organise their thinking with regards to what constitutes a lexicographical entry. These tools should be flexible enough to follow the incremental formalization process.

Associative Storage and Retrieval. Linguists also need tools to associate lemmata they have imported into the system. In particular, the Greek WordNet project associates lemmata, trying to create structures containing semantic information e.g. hyponyms, hyperonyms, homonyms, synonyms. Linguists create and browse such associations and in this way are able to find more information on specific lemmata by following specific associations. Furthermore, the system must support a personalized view of such associations: for example one linguist may relate with the association synonym the words dog and canine - indicating that dog and canine are synonyms - whilst others may not. The system should support such personalization. Associations among different data types should also be supported. E.g. the association of the lemma dog with a picture of a dog.

Classification. In the WordNet projects the notion of ontology is essential. Ontologies consist of synsets that contain lemmata or other synsets. As a consequence, they can be viewed as a tree with synsets as nodes. Synsets at different levels of the tree have different meanings. Linguists are therefore assigned the task of classifying synsets within a specific ontology. This classification however should be very flexible, since different linguists may classify synsets in different ways within the same ontology. The issue of personalization is eminent. Furthermore, linguist should be enabled to compare the ontologies they have created and find the differences, allowing this way a more co-operative working.

We identified the above patterns to be very important in the creation of language engineering tools. Of course, more patterns can be identified by examining the work and needs of linguists. The set of patterns is not complete: new semantic models can be discovered that outline the needs of linguists. Currently, most language engineering tools are monolithic: they support only one or a limited number of the aforementioned patterns. Such monolithic systems widen the gap between linguists and language engineering tool designers: while linguists discover new semantic models to capture and structure language entities, system designers fail to respond to their needs. Language engineering systems should be redesigned with the scope of capturing the new models of linguists and avoiding to reuse the existing architectures and systems. Our aim is to design an Open Language Engineering System (OLES): A language engineering system, that provides the framework in which more cognitive models - capturing more patterns as described above - can be developed and even imported within the architecture, in an effort to fill the gap between linguists and system designers.

6 Future - Parallel Plans

At present time, important knowledge and linguistic information exists in most European countries. Especially after the completion of the EuroWordNet project, scientific horizons of West European countries were greatly expanded. The success of EuroWordNet has determined the emergence of several projects that aim at the development of multilingual WordNets for the remaining European languages.

However, apart from the EuroWordNet project, no other effort has been made towards the combination and mapping across these languages. Moreover, all researchers using EuroWordNet to study and compare languages have little or no exposure to East European languages and thus, their conclusions are limited.

Our future aim is to form an relevant project for the development of a multilingual resource representing semantic relations among basic concepts of the following languages: Greek, Turkish, Bulgarian, Romanian, Serbo-Croatian and Czech.

Balkan languages have been less studied during the past years and little or not at all investigated, since they have not been deemed to have commercial impact in the short run. Furthermore, presently, information retrieval has been limited to West European data, whilst vast amounts of information written in Balkan languages still remain inaccessible. This project aims at filling this gap and making the less studied, nevertheless equally important, Balkan languages known to the entire Europe. It aims at the development a multilingual database compatible with the EuroWordNet database, enabling this way its extension with new data. The main goal of the project will be to expand the EuroWordNet project by adding to it WordNets structured for the Balkan languages.

7 Conclusions

The use of a morphosyntactic lexicon as the basis for the development of the Greek WordNet was presented in this report. The project follows the same lines as the EuroWordNet project. This way the outcome will comply with a "de facto" standard that has been set since the creation of Princeton's WordNet.

The construction of the Greek WordnNet is expected to reveal interesting aspects of linguistic phenomena as well as cross-linguistic patterns of lexicalisation.

Furthermore, it will demonstrate the feasibility of such a large scale relational lexicon for lesser studied languages, showing the way for similar projects for other Balkan languages.

References

1. Cole, R. et al.: Survey of The State of The Art in Human Language Technology. (1997). Cambridge University Press.
2. EAGLES project: Creating standards on Electronic Lexicons, Interim Reports. (1996).
3. Ide, N. and Greenstein, D.: EuroWordNet. Computers and the Humanities. **32** , (1998), Double Special Issue on EuroWordNet.
4. Egedi, D. and Martin, P.: A Freely Available Syntactic Lexicon for English. Proceedings of the Int. Workshop on Sharable Natural Language Resources, Nara, Japan, (1994).
5. Ferwell, D., Guthrie, L. and Wilks, Y.: Automatically Creating Lexical Entries for ULTRA, a Multilingual MT System. Machine Translation, **8** (1993) 127–145.
6. Grishman, R., Macleod, C. and Meyers, A.: Comlex Syntax: Building a Computational Lexicon. Project Report. (1994).
7. Knight, K. and Luk, S.: Building a Large-Scale Knowledge Base for Machine Translation. (1994).
8. Maly, K.: Compressed Tries. Communications of the ACM, **19** (1976).
9. Oostdijk, N. and deHaan, P. (Eds): Corpus-based research into language. Rodopi Publ. Amsterdam, (1994).
10. Stamison-Atmatzidi, M., Vagelatos, A., Triantopoulou, T. and Christodoulakis, D.: The Utilization of An Electronic Morphology Dictionary and a Spelling Correction System for the Teaching of Modern Greek. C.A.L.L. Journal **7** (1994) 37–49.
11. Tsalidis, C. and Orphanos, G. Word Description Languages. Proceedings of the first Workshop in Natural Language Processing, Athens, Greece, (1995) 239–253.
12. Vagelatos, A., Stamison-Atmatzidi, M., Triantopoulou, T, Farmaki, V. and Christodoulakis, D.: Analysis of Literary Style of Poet A. Sikelianos - A Computer Based Approach. "Consensus Ex Machina ?", Joint International Conference, ALLC-ACH, Sorbonne, Paris, (1994).
13. Vagelatos, A., Triantopoulou, T., Tsalidis, C. and Christodoulakis, D.: A Spelling Correction System for Modern Greek. International Journal on Artificial Intelligence & Tools. **8** (1995).
14. Vagelatos, A., Triantopoulou, T., Tsalidis, C. and Christodoulakis, D.: Utilization of a Lexicon for Spelling Correction in Modern Greek. 10th Annual Symposium on Applied Computing (SAC '95) - Special Track on Artificial Inteligence, ACM Computing Week, Nashville, Tenesse, U.S.A. (1995)
15. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, (1998).
16. WordNet: Five Papers on WordNet, International Journal of Lexicography, (1994).

This article was processed using the LTEX macro package with LLNCS style