Using a Lemmatizer to Support the Development and Validation of the Greek WordNet

Harry Kornilakis[1], Maria Grigoriadou[1], Eleni Galiotou[1,2], and Evangelos Papakitsos[1]

[1] Department of Informatics and Telecommunications, University of Athens,
Panepistimiopolis, GR-157 84, Athens, Greece
Email: harryk@di.uoa.gr, gregor@di.uoa.gr, egali@di.uoa.gr, papakitsev@vip.gr
[2] Department of Informatics, Technological Educational Institute of Athens,
Athens, Greece

Abstract. In this paper we aim to give a description of the computational tools that have been designed and implemented to support the development and validation process of the Greek WordNet, which is currently being developed in the framework of the BalkaNet project. In particular, we focus on the description of a lemmatizer for the Greek language, which has been used as the basis for a number of tools supporting the linguists in their work of developing and validating the Greek WordNet.

1 Introduction

The software infrastructure needed in view of building the Greek WordNet was developed during two consecutive projects. The DiaLexico project [3] which aimed at the construction of a lexical database with semantic relations for the Greek language and the BalkaNet project [9], which aims at the development of a multilingual lexical database with semantic relations for each of the following languages: Bulgarian, Czech, Greek, Romanian, Serbian and Turkish. The deployment of computational tools has been proved to be of major importance in the course of the aforementioned projects. The tools and resources used for the development of the monolingual Greek WordNet had to take into account the peculiarities of the Greek language, which is considered as a lesser-studied one.

In this paper we focus on the description of a lemmatizer, which has been used as the basis of a number of tools supporting the linguists in their work of extracting and processing the necessary linguistic information from dictionaries and corpora. Up to now, lemmatizers have been developed for the Greek language, mainly as tools to support specific applications, or as part of systems that support full morphological processing and require a large number of lexical resources. Examples of such systems are [5] and [8] which utilize the two-level morphology model [7] which uses a morpheme based lexicon, grammatical rules and a finite-state automaton and [6] where a lazy tagging method with functional decomposition is implemented. In our approach the lemmatizer was designed so as to be useful for a number of different tools, to require as few lexical resources as possible and to be computationally efficient.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 130–135.

2 Aspects of Greek Inflectional Morphology

Since Greek is a lesser-studied language and without the wealth of resources available for other languages, in the development of tools for the monolingual Greek WordNet we had to take into account the peculiarities of the Greek language. In this section a very brief presentation of the morphology and inflection of the Greek language that is necessary for the understanding of the rest of the paper, is given. For a more detailed description of the Greek language the reader is referred to a grammar of the Modern Greek language such as [4].

The Greek alphabet consist of 24 letters, 17 consonants (β, γ, δ, ζ, θ, κ, λ, μ, ν, ξ, π, ρ, σ, τ, φ, χ, ψ) and 7 vowels which may appear either unstressed (α, ε, η, ι, ο, υ, ω) or stressed (ά, έ, ί, ή, ό, ύ, ώ). Each word of two or more syllables has a stressed syllable that is pronounced the loudest, and in written script it is denoted by a stress mark (') over the nuclear vowel of the syllable. Each word may carry only one stress mark and according to a phonologic rule the stress may fall only upon the ultimate, penultimate or antepenultimate syllable. Word stress in Greek is *distinguishing* (e.g. *νόμος* ('nomos – law) is different from *νομός* (no'mos – administrative region). Furthermore, word stress is *moving* i.e. the stress may change its position within the inflectional paradigm of the same word. For example, the word *θάλλασα* ('*θ*alasa – sea) in the genitive plural case becomes *θαλλασών* (*θ*ala'son – of the seas).

Articles, nouns, adjectives, pronouns, verbs and participles are declinable. Nouns decline for number (singular, plural) and case (nominative, genitive, accusative, vocative), adjective decline for number, case, gender (male, female and neuter) and degree, while verbs conjugate for voice (active, passive), mood (indicative, subjunctive, imperative), tense (past, non-past), aspect (momentary, continuous), number (singular, plural) and person (1st, 2nd, 3rd) leading up to almost sixty different forms for each verb. From the above, it is easy to see that Greek is highly inflected and having to deal with each inflectional type of a word separately, would be an unnecessary burden to a linguist developing the Greek WordNet, since the citation form of each word is all that is required. Therefore, we have developed a lemmatizer for the Greek language, which can find the citation form of inflected Greek words.

3 A Lemmatizer for the Greek Language

The function of the lemmatizer is, when given as input a word in Greek, to analyze the word and to find its dictionary citation form. The lemmatizer can deal with the inflection of nouns, adjectives and verbs that do not alter their stem (which includes all derived verbs and verbs of the 2nd conjugation [4]) and can also deal with cases of irregular inflection. Furthermore it can handle stress movement. In order to achieve these, the lemmatizer keeps an amount of lexical information, which is kept in three lists: a list of words, a list of inflectional information and a list of irregular forms.

- List of words: A list containing the citation form of all the words in a dictionary.
- List of inflectional information: A list containing information about how words are inflected in Greek. Each entry in the list is of the form [*inflected_ending, citation_ending1, stress_movement1, citation_ending2, stress_movement2... citation_endingN, stress_movementN*] where each *stress_movement* is a possible ending of

the citation form of an inflected word ending in *inflected_ending*. Each *stress_movement* is a number that defines how the stress of the word moves when going from the inflected form to the citation form. Each *stress_movement* takes values between -2 and 2 that represent the following:

-2: the stress moves two syllables to the left;
-1: the stress moves one syllable to the left;
0: no stress movement;
1: the stress moves one syllable to the right;
2: the stress moves two syllables to the right.

- List of irregular forms: A list of pairs in the form [*irregular_inflected_form, citation_form]*, one pair for each irregular inflected form in the language. e.g. [*είδα, βλέπω*] where *είδα* ('i*δ*a) is an irregular form (past tense, 1st singular, indicative, active voice) of the verb *βλέπω* ('vlepo) (see).

The algorithm for lemmatizing the input word is as follows:

```
1. Search for the input word in the wordlist
If it is found
    Return the word and exit.
else
    Go to step 2
2. Search for the input word in the list of irregulars
If a pair [inflected_form, citation_form] is found
    Return citation_form and exit.
else
    Go to step 3
3. Search in the list of inflectional endings for the ending of
the input word. Find the longest possible ending that matches the
word.
If a list [inflected_ending, citation_ending1,
citation_ending2,...] is found
    Go to step 4
else
    The input word could not be lemmatized so return the input
    word and exit.
4. For each citation_ending in [citation_ending1,
citation_ending2...] do
    Remove inflected_ending from the input word
    Append citation_ending to the word
    Make the appropriate adjustment to the position of the stress
    mark on the word (See description of list of inflections above).
    Search for the new word in the wordlist.
    If it is found
        Return the word and exit.
    else
        Continue with the next citation_ending
```

```
5. If no word was found in step 4
   The input word could not be lemmatized so return the input
   word and exit.
```

4 Tools That Use the Lemmatizer

The lemmatizer has been used for three different tools whose purpose is to support the linguistic team in the development of the Greek WordNet. These tools are: A tool that counts the frequency of lemmatized word forms in text corpora, a tool that given a Greek word finds the English translation of that word and a part of speech tagger used in the annotation of corpora.

4.1 Lemmatized Word-frequency Counter

Calculating the frequency of appearance of words in corpora is useful in determining some of the base concepts. For this purpose the ECI corpus has been used. ECI is a medium-sized corpus (around 2 million words) of Modern Greek, compiled by the Universities of Edinburgh and Geneva as part of the European Corpus Initiative Multilingual Corpus. When determining base concepts it is often useful to be aware of the frequency of words in corpora, so as to avoid using as base concepts words which might be frequent in English but infrequent in Greek due to different lexical patterns between English and Greek.

The computational tool that was developed is a tool that counts the occurrences of words in corpora, in all their inflected forms. Given a number of texts in Greek the tool creates a list giving the frequency of total occurrences of each word in the texts, regardless of the inflection type in which this word appears.

In Table 1 we present an example of the results given by the word-frequency counter considering the appearances of the word *άνθρωπος* ('anθropos – man) in the ECI corpus. The frequency of each inflectional type is given separately, and in the bottom row the total occurrences of the word are given.

4.2 Translator of Words from Greek to English

The function of the word translator tool is, given a Greek word, to find the English translation of that word. The lemmatizer is a necessary component of this tool because Greek is a highly inflected language and different inflected forms of the same word may correspond to only one word form in a language with a limited inflectional system, such as English. When given a word as input this tool initially runs the lemmatizer on that word, so as to find the citation form of this word and then by looking up that word in a bilingual Greek to English dictionary we find the English translation of that word.

In the framework of WordNet development the translation is used to find the correspondence of words appearing in Greek corpora to their Inter-Lingual-Index (ILI) numbers [10]. The ILI is an unstructured list of Princeton WordNet 1.5 & 1.7 [2] synsets, with each synset in a monolingual WordNet having at least one equivalence relation with a record in this ILI. Since in the Princeton WordNet the literals of the synsets are in English, translating a Greek word to English will easily allow one to find the corresponding ILI numbers of that word.

Inflectional type	Word	Frequency
Nominative Singular	άνθρωπος	749
Genitive Singular	ανθρώπου	474
Accusative Singular	άνθρωπο	419
Vocative Singular	άνθρωπε	1
Nominative Plural	άνθρωποι	430
Genitive Plural	ανθρώπων	163
Accusative Plural	ανθρώπους	219
Total Occurrences		2455

Table 1. The count for the various inflected forms of the word "άνθρωπος"

4.3 Part of Speech Tagger

Given the lemmatizer and some information about the part of speech of words extracted from a dictionary of the Greek language, it was easy to extend the lemmatizer into a part of speech tagger for Greek texts. The wordlist was extended with part of speech information for each word, i.e. each entry in the list took the form [*word, part-of-speech1, part-of-speech2...*] allowing for each word to belong to multiple parts of speech. Therefore, once the lemmatization of a word into its citation form has been performed, we can assign a part of speech to the input word.

The extraction of the part of speech of each word was performed using the Triantafyllidis electronic dictionary of the Greek language as input and the tools developed by Galiotou et al. for the extraction of linguistic information from the definitions of electronic dictionaries [3].

This part of speech tagger is used for annotation of a Greek language corpus that is to be used as a resource for the validation of the Greek WordNet in the framework of the BalkaNet project. In particular the Greek text of George Orwell's 1984 is being annotated so as to be used for producing comparative coverage statistics for the WordNets developed as part of the project. For the rest of the languages participating in the project (except Turkish) an aligned and annotated version has already been developed as part of the Multext-East project [1], and an aligned and annotated version of the Greek text is required for acquiring reliable statistics.

5 Conclusions

In this paper, we dealt with the computational infrastructure which was developed for supporting the work of the linguists in building the Greek WordNet. In particular, we focused on the description of a lemmatizer which was used in a number of computational tools for extracting and processing linguistic information. We argued that a lemmatizer is indispensable to the processing of a highly inflected language like Greek and we described the use of the lemmatizer by other tools such a part-of-speech tagger, a word-frequency counter in corpora and a tool used for the retrieval of English translations of Greek inflected forms in a bilingual dictionary. Future work concerns the development of new tools and the enhancement of existing ones for the processing of morphosemantic information in dictionaries and corpora taking into account the particularities of the Greek language.

References

1. Erjavec, T., Ide, N., Petkevic, V., Veronis, J.: Multext-East: Multilingual Text, Tools and Corpora for Central and Eastern European Languages. Corpora Proceedings of the First TELRI European Seminar (1996) 87–98.
2. Fellbaum C. (ed.) WordNet: An Electronic Lexical Database. MIT Press (1998).
3. Galiotou E., G. Giannoulopoulou, M. Grigoriadou, A. Ralli, C. Brewster, A. Arhakis, E. Papakitsos, A. Pantelidou: Semantic Tests and Supporting Tools for the Greek WordNet, Proceedings of the NAACL Workshop on WordNet and Other Applications, Carnegie Mellon, Pittsburgh, PA, (2001) 183–185.
4. Mackridge P.: The Modern Greek Language. Oxford University Press (1985).
5. Markopoulos G.: A Two-Level Description of the Greek Noun Morphology with a Unification-Based Word Grammar. In Ralli A., Grigoriadou M., Philokyprou G., Christodoulakis D., Galiotou E. (eds.): Working Papers in NLP, Diaulos, Athens (1997).
6. Papakitsos E., Grigoriadou M., Ralli A.: Lazy Tagging with Functional Decomposition And Matrix Lexica: An Implementation in Modern Greek. Literary and Linguistic Computing, 13(4) (1998) 187–194.
7. Ralli A., Galiotou E.: Affixation in Modern Greek: a Computational Treatment. Proceedings of EURISCON '91 (1991).
8. Sgarbas K., Fakotakis N., Kokkinakis G.: A PC-KIMMO Based Morphological Description of Modern Greek.. Literary and Linguistic Computing, 10 (1995) 189–201.
9. Stamou S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D., Grigoriadou M.: BalkaNet: A Multilingual Semantic Network for Balkan Languages. Proceedings of the First International WordNet Conference, Mysore, India (2002).
10. Vossen P. (ed.): EuroWordNet: A Multilingual Database with lexical Semantic Networks. Kluwer Academic Publishers (1998).