The Software Infrastructure for the Development and Validation of the Greek Wordnet

Maria GRIGORIADOU[1], Harry KORNILAKIS[1],
Eleni GALIOTOU[1,2], Sofia STAMOU[3],
Evangelos PAPAKITSOS[1]

[1] Department of Informatics and Telecommunications,
University of Athens, Panepistimiopolis, GR-157 84, Athens, Greece
[2] Department of Informatics, Technological Educational
Institute of Athens, Athens, Greece
[3] Computer Engineering and Informatics Department,
Patras University, 26500, Greece

E-mail: {gregor, harryk, egali}@di.uoa.gr,
stamou@cti.gr, papakitsev@vip.gr

Abstract. This paper deals with the software infrastructure designed and developed to support the construction and validation of the Greek wordnet, which is currently being developed in the framework of the BalkaNet project. We describe the language resources and the tools which aim at the extraction and validation of linguistic information for wordnet development and validation. In particular, we focus on the usage of machine readable dictionaries (MRD) and corpora while taking into account the particularities of the Greek language, which play a crucial role in the aforementioned tasks. Furthermore, we emphasize on how the technical infrastructure facilitated the development of the Greek wordnet and we give an account of the methodological principles adopted to that end. Finally, we present the current status of the Greek wordnet in terms of statistical data concerning both vocabulary coverage and links representation.

1. Introduction

The software infrastructure needed in view of building the Greek wordnet was developed during two consecutive projects. The DiaLexico project [7] which aimed at the construction of a lexical database with semantic relations for the Greek language and the BalkaNet project [17], 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 particularities of the Greek language, which is considered as a lesser-studied one.

In this paper we give an overall account of this infrastructure and we distinguish between:

a) The use of MRD and the corresponding tools aiming at the extraction of the necessary linguistic information for the construction of the Greek wordnet and,

b) The usage of corpora and their processing tools employed towards the validation of wordnets (in both monolingual and multilingual perspectives), focusing on the description of a lemmatizer [9], which has been used as the backbone infrastructure for a number of peripheral tools.

Furthermore, we discuss the exploitation of the Greek version of Orwell's *1984* novel [10] for both the construction and validation of the Greek wordnet. Finally, we provide some statistical figures concerning the current status of the Greek wordnet.

2. Computational Tools Based on Machine-Readable Dictionaries

Like for all other wordnets in BalkaNet, the development of the Greek wordnet started-off with a set of concepts, common across all involved languages. The motivation behind relying on a common concept set is argued in the BalkaNet Overview paper [19] (this volume), and essentially serves conceptual overlap and compatibility issues. To that end three subsets of common terms, namely the BalkaNet Concept Set (BCS) subsets I, II and III, represented the core set of concepts to be encoded within the Greek wordnet. Selection criteria of the base concept set are again accounted for in the BalkaNet Overview paper, and are summarized as follows:

- 1 310 Subset I of EuroWordNet's (EWN) [20] base concepts, formed also BalkaNet BCS subset I.

- 3 690 most frequently occurring terms, common across all participating languages formed BCS subset II.

- BCS subset III emerged from EWN synsets common in at least five EWN wordnets, in an attempt to avoid "gaps" within BalkaNet wordnets.

Besides these common sets of terms, Greek wordnet encoded additional concepts following a merge development approach based on MRDs and corpora for Greek. These additional concepts aim at ensuring vocabulary completeness for the Greek wordnet.

In the subsequent sections, the methodology followed towards processing lexical resources for wordnet development, along with the technical infrastructure implemented towards this task are described.

Sofware Tools for Extracting Lexical Information from Dictionaries

For terminology acquisition, we processed available explanatory MRDs for Greek, the most prominent of which were:

- The explanatory dictionary of the Patakis Publishing Co. consisting of 82 021 lemmata with 67 944 definitions in a Microsoft Access 2000 format
- The "Triantafyllidis" lexicon of the Center of the Greek language consisting of 50 506 lemmata with 98 103 definitions also in a Microsoft Access 2000 format.

The software tools which were initially developed for the extraction and processing of lexical information from both lexical resources are the following:

- *The 'extraction of POS-related information' tool:* This tool extracts the definitions associated to each lemma – along with the lemma itself – either for a given part of speech (POS) or for all POS existing in the lexicon. In addition, it extracts the lemmata from the lexicon either grouped by their POS or altogether.
- *The 'extraction of linked lemmata and lemmata acting as compounds ' tool:* The electronic lexicon follows the convention that lemmata that are the feminine form of a masculine noun (e.g. *αδελφή* [sister] – *αδελφός* [brother]) or constitute the alternative form of another lemma (e.g. *αϊτόπουλο* versus *αετόπουλο* [eaglet]) are "linked" together. This tool extracts all the linked lemmata from the lexicon and also has the capability of extracting only the linked lemmata of feminine-masculine type. Similarly, this tool extracts the lemmata that can act as compounds, with or without their corresponding senses.
- *The 'synonyms and antonyms extraction' tool*: This tool extracts the synonyms and antonyms associated to each lemma. In the lexicon synonyms and antonyms are preceded by the indices *syn* and *ant* respectively.
- *The 'search for semantic relations' tool*: This tool can be used for the tracing of semantic relations such as 'role-involved', antonymic relations, 'part-of" etc. The user can search the lexicon's lemmata for certain prefixes, suffixes or morphemes in order to investigate semantic relations, for instance, acquiring an insight as to whether certain negative prefixes such as *α*- ('un-') express the relation of antonymy. There is also the choice for the user to search in the definitions of lemmata for specific expressions such as `"x is a part of y"` or `"x is a kind of y"` [2]. The tool can be used also for seeking a word or a morpheme (defined by the user), in any possible form, in the definitions of lemmata. The output contains the lemmata which contain the input word/morpheme in their definitions. The underlying assumption is that if a lemma is identified in the

definition field of another lemma, there is a possibility that a semantic relation holds between the two lemmata. Additionally, the user can look for semantic relations, based on the morphology of lemmata. The tool receives two different endings and looks for paired lemmata, which have the same or approximately the same stem and different endings. The aim of the tool is to look for possible relations such as 'role-involved' (type noun-verb indicated by endings "*-της*" and "*-ω*"), etc. in the lemmata included in the lexicon.

The contribution of the aforementioned tools in the construction of the Greek wordnet was two-fold: they assisted *(i)* the lexical selection process and *(ii)* the tracing of lexico-semantic relations between sysnets. When it comes to semantic relations encoded in Greek wordnet, the following apply:

- Lexico-semantic relations for those Greek synsets that are also represented within EWN, were inherited from EWN and subsequently validated using the aforementioned computational tools.
- Semantic relations holding between Greek synsets, not existing in EWN, were added to the Greek wordnet. Some instances of such links are: 'similar to', 'deriving from', 'also see', links that emerged after processing the definitions in the available e-dictionaries.

Since most of the semantic relations represented within BalkaNet have come from EWN and Princeton WordNet (PWN)[5] it comes naturally that the types of links encoded within all wordnets, including Greek, are similar. However, due to some technical inhibitions as well as due to the fact that cross-POS relations are enabled within Balkan wordnets, some of the PWN's relations were not used, e.g the troponymy or entailment relations holding between verbs. For these instances, their counterparts applying to nouns (i.e., hyponymy) were used.

3. Computational Tools Using Text Corpora

In this section we give a description of the computational tools, which are based on the usage of large text corpora used for the development and validation of the Greek wordnet. We especially focus on a lemmatizer for the Greek language, which is the basis for a number of other tools.

3.1. Aspects of Greek Inflectional Morphology

Since Greek is a lesser-studied language without the wealth of lexical resources existing for other languages, during the development of tools for building the Greek wordnet, we had to take into account the particularities that the Greek language exhibits. In this section, we give a brief presentation of the morphology and inflection of the Greek language, a description that is necessary for understanding the rest of the paper. For a more detailed description of the characteristics of the Greek language the reader can refer to a grammar of the Modern Greek language such as [11].

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, ή, ί, 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 phonological 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 θάλασσα ('θalasa – sea) in the genitive plural case becomes θαλασσών (θala'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 (imperfective, perfective, perfect), 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 provide the citation form of inflected Greek words.

3.2. A Lemmatizer for the Greek Language

When given as input a word in Greek, the function of the lemmatizer is to analyze the word and to find its dictionary citation form. Up to now, lemmatizers have been developed for the Greek language, mainly as tools to support specific applications, or as components of complex systems that support full morphological processing and require a large number of lexical resources. Examples of such systems are [12] and [16] which utilize the two-level morphology model, [15] which uses a morpheme based lexicon, grammatical rules and a finite-state automaton and [14] 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. 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 [11]) and can also deal with cases of irregular inflection. Furthermore it can handle stress movement. In order to achieve these, the lemmatizer maintains an amount of lexical information, which is kept in three separate lists: a list of words, a list of inflectional information and a list of irregular forms.

- *List of words*: A wordlist containing the citation form of all the words in the Greek dictionary. The list we used was based on the Triantafyllidis lexicon, enriched with some automatically generated derived forms (such as diminutives). It contains 29 782 nouns, 7 839 verbs, 12 512 adjectives and 2 067 other words.

- *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, depending on stress movement to the left or 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* ('iδa) 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 irregular words
    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, ... ] 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.
    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.
```

3.3. Computational Tools Interacting with 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 POS tagger used for corpus annotation.

The corpus we used for the development of the tools and for the extraction of linguistic information was the E.C.I. (European Corpus Initiative) Corpus. This corpus is a joint project of the Universities of Edinburgh and Geneva on behalf of the A.C.L. The Greek part of the E.C.I. has more than 2 million word-tokens, and contains approximately 89 000 word forms. These words are produced by 33 000 different lexemes, through the morphological process of inflection. The Greek part of the E.C.I. is composed of 48 files, arranged according to their subject in 11 categories. These subject categories are sports, economics, education, medicine, philosophy, astrology, law, literature, politics & sociology, science and technology. It is also organized in lists that contain the words and their individual frequencies of appearance in the text.

3.3.1. Lemmatized Word-frequency Counter

Word frequency calculation was useful in for selecting the concepts to be included in BCS subset II of the Greek wordnet. 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 E.C.I. corpus. The frequency of each inflectional type is given separately, and in the bottom row the total occurrences of the word are given.

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

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		2 455

3.3.2. Translator of Words from Greek to English

As has already been mentioned, a portion of the common set of concepts (i.e., BCS subset I) that has been encoded within individual Balkan wordnets, has been adopted from the EWN. Having determined the starting repository of concepts, it was apparent that the expand model should be followed for wordnets' development, meaning that the EWN selected concepts had to be translated to each respective language. To automate translation tasks and minimize the time and human effort overhead, we employed several available bilingual (English-to-Greek) MRDs and develop tools that would support this translation process. Such a tool is the translator of Greek words to English.

Given a Greek word, the function this tool is 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 another language with a limited inflectional system, such as English. Given a word as input, this tool initially interacts with the lemmatizer in order to find the citation form of this word. Then the English translation of that word is found by looking up that word in a bilingual Greek to English dictionary.

In the framework of wordnet development, the translation is used to find the correspondence of words appearing in Greek corpora to PWN 2.0 synsets and consequently to their BalkaNet Inter-Lingual-Index (BILI) numbers. This is possible since BILI, for the most part, follows the PWN 2.0 synsets. The contribution of the BILI in the content and structure of the Balkan wordnets is discussed in greater depth in the BalkaNet Overview paper (this volume). Since within PWN 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.

3.3.3. Part of Speech Tagger

Given the lemmatizer and some information about the POS of words extracted from a dictionary of the Greek language, it was easy to extend the lemmatizer into a POS tagger for Greek texts. The wordlist was extended with POS 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 POS. Therefore, once the lemmatization of a word into its citation form has been performed, we can assign a POS to the input word.

The extraction of the POS of each word was performed using the Triantafyllidis MRD of the Greek language as input and the tools developed by Galiotou et al. [7] for the extraction of linguistic information from the definitions of MRDs. This POS 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. Further details on the corpus used, i.e. Orwell's *1984* novel corpus are given in the next section.

3.4. Computational Tool for Concordance and Collocation Extraction

The term collocation refers to a sequence of two or more consecutive words, which often appear together in language and act as a common syntactic and semantic unit. Collocations are important in linguistic research, since their usage has established such an affinity that fluent speakers automatically associate them together. In collocations it is common that the word cluster may carry a meaning more specific or with a different nuance than could be anticipated from defining the words separately. The concordances of a word are a comprehensive index of its appearances in a corpus along with the context in which this word occurs. It is common that collocations appear in PWN synsets (e.g. *vacuum cleaner* or *talk show*), therefore in the context of wordnet development it is important to be able to locate collocations in corpora and see how common their usage is in the language so as to decide whether to include them in wordNet or not.

We have developed a tool for finding and extracting word concordances and collocations from text corpora (in our case the E.C.I. corpus was used). This tool as its input, is given a number of text files that form the corpus it will work on and a specific word, which is the target word. The tool can produces the following linguistic information regarding that corpus:

a) A list of all the concordances of the target word that appear in the text. It is possible to specify how many words to the left and to the right of the target word we want to have displayed in the concordance lines.

b) A list of the collocates of the target word sorted either alphabetically or by frequency.

4. Building a Corpus for the Validation of the Greek Wordnet

The methodology followed towards developing the Greek wordnet was common across all BalkaNet wordnets. Having decided on the common set of concepts to be represented within Greek wordnet, we followed a dual implementation approach. At the beginning we translated selected terms corresponding to PWN sysnets to Greek and we automatically imported the PWN structure, whereas later on we started developing our own synsets, paying attention so that they are linked to their corresponding sysnets in PWN. Therefore, the next step following development of the core Greek wordnet was its validation, so as to reassure the delivery of a qualitative wordnet. To that end a common methodology has again been defined for all Balkan wordnet developers. The adopted validation approach essentially implied the extensive usage of a multilingual aligned corpus for verifying the gloss completeness, the compatibility and the consistency of wordnets.

The lexical information extracted from corpora is used in the process of building and validation of the individual wordnets. In order to perform these validation tasks an aligned, annotated and lemmatized version of the Greek text of George Orwell's

1984 novel was developed. This corpus was used for producing coverage statistics for the Greek wordnet. Moreover the text, when aligned and incorporated in a multilingual parallel corpus, is used for the multilingual validation of the Balkan wordnets. Such a parallel corpus of the *1984* text has already been developed for all the participating languages in BalkaNet, except Greek and Turkish, during the Multext-East project [4].

4.1. Aligning, Annotating and Lemmatizing the Greek *1984* Corpus

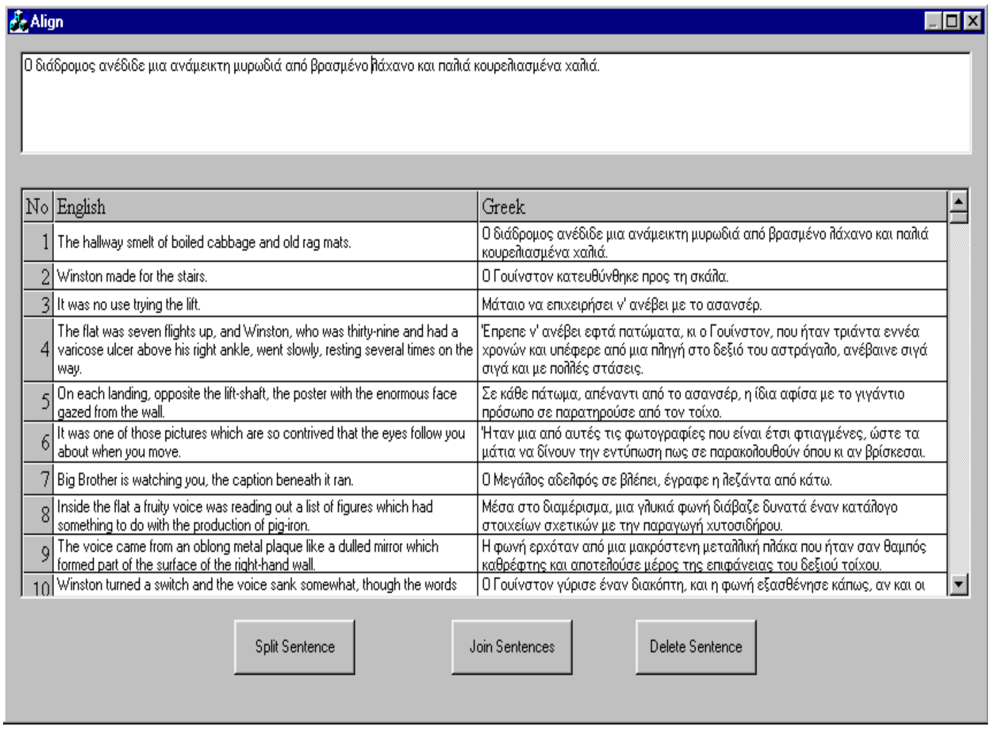
Making the Greek text of *1984* appropriate for incorporation in the multilingual corpus and therefore for wordnet's validation, initially involved the scanning of the hardcopy version of the book and the use of an Optical Character Recognition (OCR) program in order to obtain the text in machine readable form. Afterwards, it was necessary to align the text to the rest of the texts in the multilingual corpus. The final step was to annotate with morpho-lexical information and to find the citation form (lemma) of each word in the corpus.

Since English was used as a hub language we decided to align the Greek text to the English one. In this way, we were simultaneously aligning it to all the other languages. The alignment task was not trivial, since it was often the case that: *(i)* An English sentence had been translated into two Greek sentences or *(ii)* an English sentence had been left out of the Greek translation or *(iii)* a sentence of the original text was not present in the aligned corpus of the Multext-East project. This case was very common since the multilingual corpus is the set of sentences that are common to all languages. Therefore if a sentence is not present in even one of the languages it will not appear in the final multilingual corpus. Specifically, of the 6 737 sentences in the original English text only 5 466 sentences were present in the aligned multilingual version we were working with, meaning that almost 18% of the original text was missing.

Various methods for the problem of sentence alignment based mainly on machine learning have been proposed in the bibliography, for example in the works of Kay & Roescheisen [8] and Gale & Church [6]. The alignment between Greek and English sentences has also been examined in the work of Boutsis & Piperidis [1]. In the case of the *1984* corpus certain characteristics of the text made some of these methods hard to use. For example we had no previously annotated parallel corpus for training and in the English text there were no paragraph or section markers or anything else except line breaks that could be used as a delimiter. Additionally, as we mentioned before a very large part of the English text was missing making manual post-processing of the text unavoidable to a large extend.

Due to all these problems we finally opted for a more simplistic approach, which, nevertheless, would be much faster to implement. Our approach was based on a tool we have developed and that works semi-automatically. It performs an initial alignment of the text and then it offers an interface to the human editor who will correct the alignment. The initial alignment works by scanning the text for punctuation marks such as: "." "," ";" and "!", and considers these as sentence separators. Some heuristics are used in order to find the cases when these symbols don't correspond to the end of sentence. For example, when the symbol "." appears after the symbols "κ"

or "$\kappa\alpha$" ("Mr" or "Mrs") or after a single capital letter, the program assumes that this symbol is used to show abbreviation and it is not a sentence final full stop. After the first step an initial alignment of the text is achieved, but it still requires human editing, especially due to the aforementioned problems. The interface offered for this editing appears in Figure 1. The number of the sentence, the sentence in English and the sentence in Greek appear side by side. It is possible for the user to delete a sentence, to split a sentence into two sentences or to join two sentences together. Once any of those actions has been performed the numbering of the sentences is refreshed so as to reflect the new alignment between the two texts.



Ο διάδρομος ανέδιδε μια ανάμεικτη μυρωδιά από βρασμένο λάχανο και παλιά κουρελιασμένα χαλιά.

No	English	Greek
1	The hallway smelt of boiled cabbage and old rag mats.	Ο διάδρομος ανέδιδε μια ανάμεικτη μυρωδιά από βρασμένο λάχανο και παλιά κουρελιασμένα χαλιά.
2	Winston made for the stairs.	Ο Γουίνστον κατευθύνθηκε προς τη σκάλα.
3	It was no use trying the lift.	Μάταιο να επιχειρήσει ν' ανέβει με το ασανσέρ.
4	The flat was seven flights up, and Winston, who was thirty-nine and had a varicose ulcer above his right ankle, went slowly, resting several times on the way.	Έπρεπε ν' ανέβει εφτά πατώματα, κι ο Γουίνστον, που ήταν τριάντα εννέα χρονών και υπέφερε από μια πληγή στο δεξιό του αστράγαλο, ανέβαινε σιγά σιγά και με πολλές στάσεις.
5	On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall.	Σε κάθε πάτωμα, απέναντι από το ασανσέρ, η ίδια αφίσα με το γιγάντιο πρόσωπο σε παρατηρούσε από τον τοίχο.
6	It was one of those pictures which are so contrived that the eyes follow you about when you move.	Ήταν μια από αυτές τις φωτογραφίες που είναι έτσι φτιαγμένες, ώστε τα μάτια να δίνουν την εντύπωση πως σε παρακολουθούν όπου κι αν βρίσκεσαι.
7	Big Brother is watching you, the caption beneath it ran.	Ο Μεγάλος αδελφός σε βλέπει, έγραφε η λεζάντα από κάτω.
8	Inside the flat a fruity voice was reading out a list of figures which had something to do with the production of pig-iron.	Μέσα στο διαμέρισμα, μια γλυκιά φωνή διάβαζε δυνατά έναν κατάλογο στοιχείων σχετικών με την παραγωγή χυτοσιδήρου.
9	The voice came from an oblong metal plaque like a dulled mirror which formed part of the surface of the right-hand wall.	Η φωνή ερχόταν από μια μακρόστενη μεταλλική πλάκα που ήταν σαν θαμπός καθρέφτης και αποτελούσε μέρος της επιφάνειας του δεξιού τοίχου.
10	Winston turned a switch and the voice sank somewhat, though the words	Ο Γουίνστον γύρισε έναν διακόπτη, και η φωνή εξασθένησε κάπως, αν και οι

Fig. 1. The sentence alignment tool.

After the Greek text had been aligned to the multilingual text, it was necessary to annotate the words in the text with their grammatical attributes and to lemmatize them i.e. find the citation form for each word. Even though POS taggers have been proposed for the Greek language in papers such as [3] and [13], these methods are mainly based on machine learning and require an annotated training corpus in order to work, a requirement which in our case was not available. Our approach was to use a lemmatizer for the Greek language whose function is, when given as input a word in Greek, to analyze the word and to find its dictionary citation form. A description of the lemmatizer is given in section 3.

4.2. Greek Corpus Characteristics

The annotated text follows the specification given in the Multext-East project. In Table 2 we give the attributes for each part of speech used in the Greek corpus along with the number of words that belong to that POS in the corpus. A sample sentence from the corpus, as it has been annotated for Greek, appears in Figure 2.

Table 2. Parts of speech with their attributes and their frequency in the *1984* corpus

POS	Attributes	Appearances
Noun	Type Gender Number	17 047
Verb	Mood Tense Person Number Voice Aspect	14 985
Adjective	Degree Gender Number	6 394
Pronoun	Type	7 542
Article	Type	11 329
Adposition	Type	6 298
Conjunction	Type	5 123
Numeral	Type	1 041
Particle	Type	4 926
Interjection	–	9
Abbreviation	–	21

```
⟨tu id=”Ozz.3751”⟩
⟨seg lang=”gr”⟩⟨s⟩
⟨w lemma=”ακούω”
ana=”V-is3s-p——e”⟩ Ακούστηκε⟨/w⟩
⟨w lemma=”πάλι” ana=”R-p”⟩πάλι⟨/w⟩
⟨w lemma=”ένας” ana=”Ti”⟩ένας⟨/w⟩
⟨w lemma=”πάταγος” ana=”Ncms”⟩πάταγος⟨/w⟩
⟨c⟩.⟨/c⟩⟨/s⟩⟨/seg⟩
⟨/tu⟩
```

Fig. 2. Sample sentence of the Greek *1984* corpus.

4.3. Utilizing the Corpus towards Wordnet Validation

Having successfully developed, aligned and processed the Greek part of the corpus, it was used for the validation of the Greek wordnet. Validation was essentially two-fold, including the monolingual validation i.e. testing the quality of each individual wordnet, and the multilingual validation i.e. testing the consistency of relations holding among words across the various Balkan wordnets. Monolingual validation is performed by producing coverage statistics of the corpus by the wordnet itself. To perform that, we find all the lemmata of the corpus and the we check how many of them are also found in Greek wordnet. In this way we can locate words missing from the wordnet and enrich it.

The idea behind the multilingual validation is to use the parallel corpora in order to find the relations among words in the various languages. By using tools that can automatically construct translation lexicons from annotated parallel corpora [18] it is possible to create bilingual wordlists for each of the pairs of languages. Once such wordlists are available they will be used for the multilingual validation of BalkaNet by seeing if the relations between words that appear on these multilingual wordlists agree with the relations between the same words in the wordnets of the two languages.

5. Current Status of Greek Wordnet

The tools described so far have been designed and implemented in such a way so as to support the development and the validation of the Greek wordnet. In this section we outline the main methodological approaches followed while developing Greek wordnet and we provide some statistical data concerning its vocabulary coverage and the representation of its relations.

Greek wordnet, being a part of the BalkaNet conceptual warehouse, has been implemented having the same methodological considerations as the rest of the Balkan wordnets. These concern the representation of a common set of concepts across all languages, i.e., BCS subsets I, II and III, the linking of monolingual synsets to their PWN 2.0 corresponding synsets and the adoption of EWN's lexico-semantic relations, of which hyponymy, antonymy and meronymy were to be used in all Balkan wordnets. Based on the above specifications, and via the usage of a scalable tool-kit we have developed, we managed to come up with a core wordnet for Greek. This core wordnet being in perfect alignment with the rest Balkan wordnets, was comprised of approximately 8 500 synsets. Following on, enrichment of the wordnet took place based on language-specific properties and monolingual lexical resources. Enrichment was vital for two reasons. To tackle wordnet lexical gaps, i.e., empty nodes occurring when English synsets had no Greek lexicalized counterpart and to reflect lexico-semantic relations holding between Greek concepts. During enrichment, monolingual lexical resources have been processed in ways described above and selected lexical elements were extracted out of them in order to form the additional Greek wordnet's synsets. These new synsets were linked to their PWN 2.0 counterparts via the structure of BalkaNet's Inter-Lingual-Index (BILI), which helped reassure a level of consistency between Balkan wordnets. Synsets were also linked to other synsets in the Greek wordnet through the usage of semantic relations.

The status of the Greek wordnet as of this writing is illustrated in Table 3. Specifically, the total number of synsets, literals and their ratio are given. Moreover, the total number of language internal relations between Greek wordnet, as well as the average ratio of links per synset. Finally, the table illustrates the number of non-lexicalized concepts, and the total number of glosses encoded for Greek wordnet synsets. In table 4 we provide the number of synsets for each BCS subset along with the POS distribution of all Greek wordnet synsets. The numbers of semantic relations in the Greek wordnet so far, for each type of relation is given in Table 5.

Table 3. Statistical Data on Greek Wordnet

Total Number of Synsets	18 677
Literals	24 811
Ratio Literals/Synsets	1.33
Relations	24 582
Ratio Relations/Synsets	1.33
Non-lexicalized Concepts	46
Glosses	18 649

Table 4. POS and BCS Distribution for Greek Wordnet

BCS Subset	**Number**
BCS I	1 218
BCS II	3 462
BCS III	3 826
Part of Speech	**Number**
Nouns	14 480
Verbs	3 538
Adjectives	635

Table 5. Semantic Relations in the Greek Wordnet

Relation Type	**Number**
Also See	210
Be In State	143
Verb Group	424
Derived	64
Holo-member	1 324
Holo-part	2 708
Holo-portion	162
Hypernym	18 521
Holo-substance	57
Causes	76
Near Antonym	693
Similar To	46
Subevent	132
Antonym	22
Total	**24 582**

Finally, a critical issue while developing the Greek wordnet, besides the actual implementation per se, was wordnet's validation control. Validation tasks were focused on both semantic and syntactic validation. The approaches against semantic validation (both monolingual and multilingual) have been outlined in a previous section of the paper. Syntactic validation was a task performed by each member of the Balkan team individually and concerned mainly checking the following within wordnets:

- Literals and synsets' correct spelling.
- Validity of the XML representation scheme.
- Denotation of the empty nodes using the ⟨NL⟩ . . . ⟨/NL⟩ tag.
- Validity of the representation of IDs and removal of duplicate IDs.
- Ensure that there are no empty tags.
- Removal of duplicates, be it either synsets, literals or relations.

6. Conclusions

In this paper, we dealt with the computational infrastructure which was developed developed in support of building the Greek wordnet. Specifically we described the lexical resources and the software tools which were developed for the extraction and processing of the necessary linguistic information, taking into account the particularities of the Greek language. We focused on the description of a lemmatizer which was used in a number of computational tools for extracting and processing lexical information. We argued that a lemmatizer is indispensable to the processing of a highly inflected language, like Greek, and we described the utilization of the lemmatizer in other tools such a POS tagger, a word-frequency counter and a tool used for the retrieval of English translations of Greek inflected forms in a bilingual dictionary. We also described the process of alignment, annotation and lemmatization of the Greek version of Orwell's *1984* novel and its exploitation for the validation of monolingual and multilingual wordnets. Finally, we have outlined the main steps adopted against the implementation and validation of the Greek wordnet and provided a statistical overview of its current status.

Future work concerns the development of new tools and the enhancement of existing ones for the processing of morphosemantic information in both dictionaries and corpora keeping in mind the particularities of the Greek language.

References

[1] BOUTSIS, S., PIPERIDIS, S., *Automatic Extraction of Bilingual Lexical Equivalences from Parallel Corpora*, in *Proc. Multilinguality in Software Industry /ECAI*, August 27–31, Budapest, Hungary, 1996.

[2] CHARCHARIDOU, A., GRIGORIADOU, M., *Meronyms Computational Extraction from Dictionary Definitions in a Semi-automatic Way*, in *Workshop on International Proofing Tools and Language Technologies*, Patras, Greece, July 2004.

[3] DERMATAS, E., KOKKINAKIS, G., *Automatic Stochastic Tagging of Natural Language Texts*, Computational Linguistics, **21**, 2, 1995.

[4] ERJAVEC, T., IDE, N., PETKEVIC, V., VERONIS, J., *Multext-East: Multilingual Text, Tools and Corpora for Central and Eastern European Languages*, in *Corpora Proceedings of the First TELRI European Seminar*, 87–98, 1996.

[5] FELLBAUM, C. (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[6] GALE, W., CHURCH, K., *A Program for Aligning Sentences in Bilingual Corpora*, Computational Linguistics, **19**, 1993.

[7] GALIOTOU, E., GIANNOULOPOULOU, G., GRIGORIADOU, M., RALLI, A., BREWSTER, C., ARHAKIS, A., PAPAKITSOS, E., PANTELIDOU, A., *Semantic Tests and Supporting Tools for the Greek Wordnet*, in *Proceedings of the NAACL Workshop on WordNet and Other Applications*, Carnegie Mellon, Pittsburgh, PA, 183–185, 2001.

[8] KAY, M., ROESCHEISEN, M., *Text-translation Alignment*, Computational Linguistics, **19**, 121–142, 1993.

[9] KORNILAKIS, H., GRIGORIADOU, M., GALIOTOU, E., PAPAKITSOS, E., *Using a Lemmatizer to Support the Development and Validation of the Greek Wordnet*, in *Second Global Wordnet Conference*, Brno, Czech Republic, January 2004.

[10] KORNILAKIS, H., GRIGORIADOU, M., GALIOTOU E., PAPAKITSOS E., *Aligning, Annotating and Lemmatizing a Corpus for the Validation of Balkan Wordnets*, in *Workshop on Balkan Language Resources and Tools,* Thessaloniki, Greece, November 2003.

[11] MACKRIDGE, P., *The Modern Greek Language*, Oxford University Press, 1985.

[12] 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*, Diavlos, Athens, 1997.

[13] PAPAGEORGIOU, H., PROKOPIDIS, P., GIOULI, V., PIPERIDIS, S., *A Unified POS Tagging Architecture and its Application to Greek*, *Proceedings of Second International Conference on Language Resources and Evaluation-LREC2000*, 1455–1462, Athens, Greece, 2000.

[14] 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, 187–194, 1998.

[15] RALLI, A., GALIOTOU, E., *A Morphological Processor for Modern Greek*, in *Proceedings of EACL-87,* Copenhagen, 26–31, 1987.

[16] SGARBAS, K., FAKOTAKIS, N., KOKKINAKIS, G., *A PC-KIMMO Based Morphological Description of Modern Greek*, Literary and Linguistic Computing, **10**, 189–201, 1995.

[17] STAMOU, S., OFLAZER, K., PALA, K., CHRISTOUDOULAKIS, D., CRISTEA, D., TUFIŞ, D., KOEVA, S., TOTKOV, G., DUTOIT, D., GRIGORIADOU, M., *BalkaNet: A Multilingual Semantic Network for Balkan Languages*, in *Proceedings of the First International WordNet Conference*, Mysore, India, 2002.

[18] TUFIŞ, D., BARBU, A.M., *Automatic Construction of Translation Lexicons*, in *Proceedings of the WSES and IEEE International Conference on Multimedia, Internet, Video Technologies*, 2181–2186, Malta, September 1–6, 2001.

[19] TUFIŞ, D., CRISTEA, D., STAMOU S., *BalkaNet: Aims, Methods, Results and Perspectives. A General Overiew*, Romanian Journal of Information Science and Technology, **7**, nos. 1–2, 2004 (in this volume).

[20] VOSSEN, P. (ed.), *EuroWordNet: A Multilingual Database with lexical Semantic Networks*, Kluwer Academic Publishers, 1998.