ALIGNING, ANNOTATING AND LEMMATIZING A CORPUS FOR THE VALIDATION OF BALKAN WORDNETS

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

[1] Department of Informatics and Telecommunications, University of Athens,
Panepistimiopolis, GR-157 84, Athens, Greece
{harryk, gregor, egali}@di.uoa.gr, papakitsev@vip.gr

[2] Department of Informatics, Technological Educational Institute of Athens, Athens, Greece

Keywords: WordNet validation, sentence alignment, corpus annotation.

Abstract

In this paper we discuss the usage of corpora in the validation of WordNets and we present the exploitation of the Greek version of George Orwell´s Nineteen *Eighty-Four* for the construction and validation of the Greek WordNet, which is currently under development in the framework of the BalkaNet project. In particular, we focus on the description of tools that were developed and used for the alignment, the annotation and the lemmatization of the corpus.

1. Introduction

The Greek WordNet has been developed in the course of two national and European funded consecutive projects:

a. The DiaLexico project (Galiotou et al. 2001) which aimed at the construction of a lexical database with semantic relations for Greek.

b. The BalkaNet project (Stamou et al. 2002) which aims at the development of a multilingual database with WordNets for the Balkan languages (Bulgarian, Czech, Greek, Romanian, Serbian and Turkish) following the principles of WordNet (Fellbaum, 1998) and EuroWordNet (Vossen 1998).

In the course of building the Greek WordNet, we have built a number of computational tools in order to extract the necessary linguistic information from electronic dictionaries and corpora.

The linguistic information extracted from corpora is used in the process of building and validation of the individual WordNets. In particular, the annotation and lemmatized version of the Greek text of George Orwell´s *Nineteen Eighty-Four* is used for producing coverage statistics for the Greek WordNets developed as part of the BalkaNet project. Moreover this 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 *Nineteen Eighty-Four* text has already been developed for all the participating languages in BalkaNet, except Greek and Turkish, during the Multext-East project (Erjavec et al., 1996 & 1998). In order to perform these validation tasks an aligned, annotated and lemmatized version of the Greek text is being developed.

2. The Multilingual *1984* Corpus

As part of the Multext-East project, the *Nineteen Eighty-Four* text was aligned, annotated and lemmatized for the following languages: Romanian, Slovene, Czech, Bulgarian, Estonian, Hungarian and English. Later on the text was also aligned, annotated and lemmatized for Serbian. Therefore the only languages of countries participating in BalkaNet which are not also part of the multilingual *Nineteen Eighty-Four* are Greek and Turkish.

For the annotation of the text, a standardized specification for the description of the morpho-lexical information of words was proposed (Tufis et al., 1998) in the framework of the Multext-East project. The morpho-lexical information is provided as a string, using a linear, term-like encoding. In this notation, the position in a string of characters corresponds to an attribute, and specific characters in each position indicate the value for the corresponding attribute. That is, the positions in a string of characters are numbered 0, 1, 2, etc., and are used in the following way:

- The character at position 0 encodes part-of-speech;
- Each character at position 1, 2, n, encodes the value of one attribute (person, gender, number, etc.), using a one-character code.
- If an attribute does not apply, the corresponding position in the string contains the special marker ‘-'.

For example, the string *"Ncns"* stands for:

Part-of-speech: **N**oun
Type: **c**ommon
Gender: **n**euter
Number: **s**ingular

Each sentence in the multilingual corpus is assigned a sentence number which uniquely identifies it. Sentences with the same number are common for all languages. An example of such a sentence appears in Figure 1. The sentence with number 3751 appears in English, Romanian and Czech. The annotation on the text is done with XML and for each word its dictionary citation form ("lemma" attribute) and its morpho-lexical information ("ana" attribute) is given. As it can be seen in the figure the English word *"crash"* is assigned the grammatical information *"Ncns"* which, as mentioned, means that it is a common neuter, singular noun.

```
<tu id="Ozz.3751">
<seg lang="en"><s id="Oen.2.10.33.8">
<w lemma="there" ana="Pt3">There</w>
<w lemma="be" ana="Vmis3s">was</w>
<w lemma="another" ana="Dg--s">another</w>
<w lemma="crash" ana="Ncns">crash</w>
<c>.</c></s></seg>

<seg lang="ro"><s id="Oro.2.10.70.6">
<w lemma="sine" ana="Px3--a--------w">Se</w>
<w lemma="auzi" ana="Vmis3s">auzi</w>
<w lemma="un" ana="Tifsr">o</w>
<w lemma="nou"
ana="Afpfsrn">nou&abreve;</w>
<w lemma="bufnitur&abreve;"
ana="Ncfsrn">bufnitur&abreve;</w>
<c>.</c></s></seg>

<seg lang="cs"><s id="Ocs.2.10.33.8">
<w lemma="zazn&iacute;t"
ana="Vmps-sfan----n">Zazn&ecaron;la</w>
<w lemma="dal&scaron;&iacute;"
ana="Afpfsn---c">dal&scaron;&iacute;</w>
<w lemma="r&aacute;na"
ana="Ncfsn">r&aacute;na</w>
<c>.</c></s></seg>
</tu>
```

Figure 1: An annotated, aligned and lemmatized sentence for English, Romanian and Czech taken from the Multext-East project.

3. Building the Greek *1984* Corpus

Making the Greek text of *Nineteen Eighty-Four* 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 is to annotate with morpho-lexical information and find the citation form (lemma) of each word in the corpus.

3.1 Sentence Alignment

The purpose of the sentence alignment process is to take each sentence in the Greek text and find which is the corresponding sentence in the English text. By aligning to the English text, we are simultaneously aligning to all the other languages, since English in Multext-East was used as a hub language.

The alignment task is not trivial, since it is often the case that one of the following problems exists:

1. An English sentence has been translated into two Greek sentences e.g. "*Winston found and handed over two creased and filthy notes, which Parsons entered in a small notebook, in the neat handwriting of the illiterate.*" is translated as *"Ο Γουίνστον έβγαλε κι έδωσε δύο τσαλακωμένα και βρόμικα χαρτονομίσματα. Ο Πάρσονς, με το καθαρό γράψιμο του αγράμματου, σημείωσε το ποσόν σ' ένα μικρό σημειωματάριο."*
2. Two or more English sentences have been translated into one Greek sentence. e.g. *"It was partly the unusual geography of the room that had suggested to him the thing that he was now about to do. But it had also been suggested by the book that he had just taken out of the drawer."* is translated with the single sentence *"Λίγο αυτή η ασυνήθιστη γεωγραφία του δωματίου, λίγο τούτο το τετράδιο που μόλις είχε βγάλει από το συρτάρι, του είχαν υποβάλει την ιδέα να κάνει ό,τι ετοιμαζόταν να κάνει τώρα."*
3. An English sentence has been left out of the Greek translation.
4. A sentence of the original text is not present in the aligned corpus of the Multext-East project. This case is very common since the multilingual corpus is the set of sentences that are common for all languages. Therefore if a sentence was not present in even one of the languages it will not appear in the final multilingual corpus. Specifically, of the 6737 sentences in the original English text only 5466 sentences were present in the aligned multilingual version we were working with, meaning that almost 18% of the original text was missing.

The 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 (1993) and Gale & Church (1993). The alignment between Greek and English sentences has also been examined in (Boutsis & Piperidis, 1996). In the case of the 1984 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 necessary 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 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 “κα” (“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.

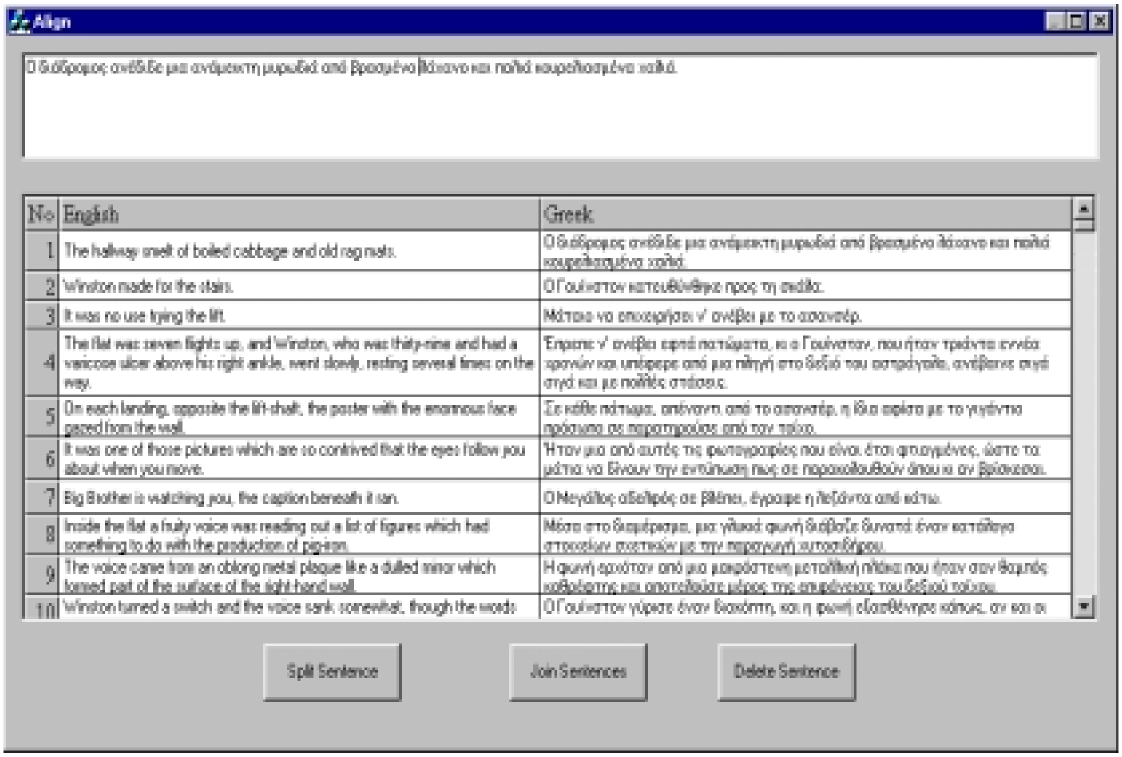
After the first step an initial alignment of the text is achieved, but it still requires human editing, especially due to the aforementioned problems. The interface offered for this editing appears in Figure 2. The number of the sentence, the sentence in English and the sentence in Greek appear side 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.

3.2. Annotation and Lemmatization

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. for each word find its citation form.

Part of Speech taggers have been proposed for the Greek language in papers such as (Dermatas & Kokkinakis 1995), (Papageorgiou et al, 2000) and (Petasis et al, 1999). However these methods are mainly based on machine learning and require an annotated training corpus in order to work, 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 is to analyze the word and to find its dictionary citation form. The lemmatizer can deal with the inflection of nouns, adjectives and with the conjugation of verbs that do not alter their stem (which includes all derived verbs and verbs of the 2nd conjugation (Mackridge, 1985)) and can also deal with cases of irregular inflection. Furthermore it can handle stress movement. In order to achieve these, the lemmatizer should keep 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 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 29782 nouns, 7839 verbs, 12512 adjectives and 2067 other words.



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.	Μάταια 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 εννιά 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 της 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

Figure 2: The sentence alignment tool

Language	Greek	English	Bulgarian	Romanian	Czech
Tokens	93299	118102	101173	118063	100358
Words	81316	103997	86020	101508	79862
Distinct Words	12972	9745	16348	15225	19115
Distinct Lemmas	6375	7260	8517	7433	9161

Table 1: Characteristics of the multilingual corpus for the various languages

- List of inflectional information: A list containing information about how words are inflected in the Greek language. Each entry in the list is of the form
  [*inflected_ending,*
  citation_ending1,
  stress_movement1,
  lexical_information1,
  citation_ending2,
  stress_movement2,
  lexical_information2,
  …
  citation_endingN,
  stress_movementN,
  lexical_informationN]
  where each *citation_ending* is a possible ending of the citation form of an inflected word ending in *inflected_ending*. S*tress_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 movement of the word stress one or two syllables to the left or to the right.
  Lexical_information is the morpho-lexical information of the inflected

word, encoded with the Multext-east specifications.

- List of irregulars: A list of triplets in the form *[irregular_inflected_form, citation_form, lexical_information]*, one pair for each irregular inflected form in the language. e.g. [*είδα, βλέπω, V-is1s-a------e*] where *είδα* ('iδa) is an irregular form (indicative, past tense, 1st singular, active voice, perfective) of the verb *βλέπω* ('vlepo) (see). The list of irregulars we used was not extensive and contained around 400 of the most commonly used irregular forms of Greek.

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 triplet of the form [inflected_form, citation_form, lexical_information], is found
    Return citation_form and lexical_information 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 matching list 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 the corresponding lexical_information 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. Greek Corpus Characteristics

In table 1 we present the characteristics of the Greek text of *Nineteen Eighty-Four* is comparison to the same data for the rest of the languages which are common in both Multext-East and BalkaNet. Data for the language except Greek were taken from (Dimitrova et al., 1998). It can be seen that the numbers are comparable for all languages.

The annotated text follows the specification given in the Multext-East project. In table 2 we give the attributes for each part of speech and the number of words that belong to that part of speech in the corpus. A sample sentence from the corpus, as it has been annotated for Greek, appears in Figure 3. In fact, it is the sentence that was given in Figure 1 for English, Romanian and Czech.

```
<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>
```

Figure 3: Sample sentence of the Greek corpus.

POS	Attributes	Appearances
Noun	Type Gender Number	17047
Verb	Mood Tense Person Number Voice Aspect	14985
Adjective	Degree Gender Number	6394
Pronoun	Type	7542
Article	Type	11329
Adposition	Type	6298
Conjunction	Type	5123
Numeral	Type	1041
Particle	Type	4926
Interjection	-	9
Abbreviation	-	21

Table 2: The parts of speech that can be found in the corpus, their attributes and their frequency.

5. Using the Corpus for WordNet Validation

Once the corpus has been created, the next step is to use it for the validation of WordNet. It can support both the monolingual validation i.e. testing the quality of each individual WordNet, and the multilingual validation i.e. testing the relations among words across the various Balkan WordNets.

Monolingual validation is performed by producing coverage statistics of the corpus by the WordNet. To perform this all the lemmata are found in the corpus and then we check to see how many of them can be found in the Greek WordNet. This way we can find 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 (Tufis & Barbu, 2001) 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.

6. Conclusions

In this paper we describe the usage of corpora in the process of validation of the WordNet developed during the BalkaNet project. In particular, we have presented the current status of our work which consists in the alignment, annotation and lemmatization of the corpus and the development of tools for the production of coverage comparative statistics for the WordNets developed during the BalkaNet project. Future work aims at comparing our results with results obtained by other Part of Speech taggers for the Greek language in order to test the quality of the annotation and the lemmatization.

References

Boutsis, S., and S. Piperidis. (1996). Automatic Extraction of Bilingual Lexical Equivalences from Parallel Corpora. In Proc. Multilinguality in Software Industry /ECAI, 27-31.12 August, Budapest, Hungary.

Dermatas E., Kokkinakis G. (1995) *Automatic Stochastic Tagging of Natural Language Texts.* Computational Linguistics 21:2

Dimitrova, L., Erjavec, T., Ide, N. Kaalep, H., Petkevic, V., Tufis, D. (1998) Multext-East: Parallel and Comparable Corpora for Six Central and Eastern European Languages. *Proceedings of ACL/COLING98,* Montreal, 315-19.

Erjavec, T. and Ide, N. (1998) The MULTEXT-EAST Corpus, *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, pp. 971-974

Erjavec, T., Ide N., Petkevic, V., Veronis, J. (1996) Multext-East: Multilingual Text, Tools and Corpora for Central and Eastern European Languages, *Proceedings of the First TELRI European Seminar*, pp.87-98.

Fellbaum C. (ed.) (1998) WordNet: An Electronic Lexical Database. MIT Press

Gale W.; Church K. (1993) *A Program for Aligning Sentences in Bilingual Corpora.* Computational Linguistics 19.

Galiotou E., G. Giannoulopoulou, M. Grigoriadou, A. Ralli, C. Brewster, A. Arhakis, E. Papakitsos, A. Pantelidou (2001) Semantic Tests and Supporting Tools for the Greek WordNet, *Proceedings of the NAACL Workshop on WordNet and Other Applications*, Carnegie Mellon University, Pittsburgh, PA, pp. 183-185.

Kay M., Roescheisen M. (1993). Text-translation Alignment. Computational Linguistics 19. 121-142.

Mackridge P. (1985) The Modern Greek Language. Oxford University Press.

Papageorgiou, H., Prokopidis, P., Giouli, V., Piperidis, S. (2000) A Unified POS Tagging Architecture and its Application to Greek. *Proceedings of Second International Conference on Language Resources and Evaluation-LREC2000*, 31 May- 2 June 2000, Athens, Greece, 1455-1462.

Petasis G., Paliouras G., Karkaletsis V., Spyropoulos C.D., Androutsopoulos I. (1999): Resolving Part-of-Speech Ambiguity in the Greek Language Using Learning Techniques. In Fakotakis, N. et al. (Eds.), *Machine Learning in Human Language Technology*. Proceedings of the ACAI Workshop 29-34, Chania, Greece, 1999.

Stamou, S., Oflazer, K., Pala, K., Christodoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, D., Dutoit, D., Grigoriadou, M., (2002) BalkaNet: A Multilingual Semantic Network for Balkan Languages, *Proceedings of the First Intrenational WordNet Conference*, Mysore

Tufis D., Barbu A.M. (2001) Automatic Construction of Translation Lexicons, in *Proceedings of the WSES and IEEE International Conference on Multimedia, Internet, Video Technologies*, Malta, 1-6 September, 2001, pp. 2181-2186.

Tufis, D., Ide, N., Erjavec, T. (1998) Standardized Specifications, Development and Assessment of Large Morpho-syntactic Resources for Six Central and Eastern European Languages. *First International Language Resources and Evaluation Conference,* Granada, Spain, 233-40.

Vossen P. (ed.) (1998) EuroWordNet: A Multilingual Database with lexical Semantic Networks. Kluwer Academic Publishers