Proc. of IS 2005, the 8th International Multiconference on in Information Society, 11-17 October 2005, Ljubljana pp. 199-202
Conference on Data Mining and Data Warehouses (SiKDD 2005)

LEARNING POS TAGGING FROM A TAGGED MACEDONIAN TEXT CORPUS

Viktor Vojnovski[1], *Sašo Džeroski*[2], *Tomaž Erjavec*[2]
[1]Institute of Informatics, Faculty of Natural Sciences and Mathematics
Arhimedova 5, 1000 Skopje, Macedonia
[2]Department of Knowledge Technologies, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: sparks@on.net.mk,{saso.dzeroski,tomaz.erjavec}@ijs.si

ABSTRACT

This paper presents several new linguistic resources for the Macedonian language, in particular a language corpus consisting of the digitized and annotated Orwell's "1984" in the Macedonian translation. The produced resources (morphosyntactic specification, lexicon, and corpus) are compatible with the multilingual MULTEXT-East data set. The paper presents the digitisation, up-conversion, alignment, and annotation of the corpus, and then discusses an initial experiment in training and evaluating a Part-of-Speech tagger for the Macedonian language on the produced corpus.

1 INTRODUCTION

Abundant is not the word one would use to describe the amount of linguistic resources available for the Macedonian language. The international linguistic community realized the advantages of a computer based approach long ago, and now computational linguistics is one of the most prominent directions in linguistics. It has to be said that the majority of work done focuses on English, while research in computational linguistics in the context of other languages is much more contained.

The first project to encompass linguistic resources for East-European languages was MULTEXT-East. The EU funded project ended in 1997, but resources for new languages are still being developed [1]; currently MULTEXT-East contains resources for Bulgarian, Czech, Estonian, Hungarian, Romanian, Russian, Slovene, Croatian and Serbian, as well as for English, the "hub" language of the project.

The work described in this paper builds on work by Zdravkova [2], which makes the first contribution to the development of Macedonian morpholexical resources according to the guidelines of MULTEXT-East. The development of the resources was made in three stages. First, morphosyntactic specifications were developed for the Macedonian language. These define the so-called morphosyntactic descriptions (MSDs), which express word-class syntactic information. The second stage was building word-form lexica, which cover the lexical stock of the corpus. Currently, the Macedonian MSD system is fully defined, and an initial attempt at creating a lexical collection was made. In this paper we will build upon this work and use the MSDs to annotate the novel "1984" by G. Orwell, therefore obtaining yet another piece of the MULTEXT-East puzzle.

While PoS tagging is not a new research topic, it is a new field as far as East-European languages are concerned. These languages typically have quite different properties, in particular a much richer word inflection. An even greater problem is the lack of training and testing data, i.e., pre-annotated corpora. In this paper, we will present the first PoS tagging learning and evaluation study ever made on a Macedonian corpus.

The paper is organized as follows. In Section 2, the process of digitalization of Orwell's "1984" is described, giving as a result a version of the book in a standardized format. What follows is the task of tokenizing the text into contextual units: paragraphs, sentences, and words, along with encoding in TEI format, which is the topic of Section 3. Section 4 presents the work done on aligning the sentences of the Macedonian version of 1984 with the English one. Next, Section 5 tackles the problem of learning and evaluation of a PoS tagger over the newly created corpus. Finally, Section 6 concludes the paper by discussing the results and proposes directions for further work.

2 DIGITALIZATION OF ORWELL'S "1984"

The Macedonian translation of Orwell's "1984" is relatively new [3], but nevertheless it suffers from grammatical errors in the translation, as well as many errors introduced during print. Furthermore, no digital version of the text was available to the authors, despite the fact that it was digitally typeset. So, we were facing the dilemma of typing the text anew, or using OCR methods of converting it to digital form. While typing the text from scratch could have helped solve many of the errors so prominent in the book, it was our opinion that it would undoubtedly introduce many new ones, and would surely take too much time.

Error correction

The book was scanned in a Microsoft Word format using ABBYY FineReader. Choosing the OCR method presented us with many challenges. Incorrectly scanned characters were the first problem we sought to solve. The scanner recognized many Cyrillic characters as Latin ones, namely those whose glyph is shaped the same as a Latin one, and this had to be subsequently corrected.

As the Macedonian spellchecker that comes with Microsoft Office uses a very small wordlist, a list of the most common 300,000 Macedonian words found on Macedonian web pages was used [4]. More that 2000 words were corrected using the spellchecker. The next problem was the recognition of certain typographical characters, such as the dash sign, which was incorrectly interpreted as a minus sign. All of the abovementioned errors, along with a plethora of others, were corrected, resulting with a proofread version of the text.

Technical details

Conversion from Microsoft Word to XML was done using the program UpCast. The output from UpCast is fed through several XSLT conversion scripts, in order to get a TEI encoded version of the text, where the smallest unit of division is a paragraph. It has to be noted that during the process, a whole framework of programs in Perl, driven by Makefiles, was written. Therefore, the entire process is automatic, and all the documents mentioned in the paper can be generated from the initial scanned Word document.

3 TOKENIZING AND ENCODING

Tokenization

The corpus was tokenized using the Perl program `mltokenizer`, which was written during the development of the tool `totale` [5]. It works by splitting the text into tokens (according to language dependent resources, such as lists of abbreviations) and assigning a type to each token. The types distinguish not only words form punctuation, but also mark digits, abbreviations, clitics etc. The tokenizer also marks ends of paragraphs and sentences. Figure 1 is a sample of the tokenized text.

```
TAG       <div type="chapter" n="1" id="Omk.1.1">
TAG       <head>
TOK       I
TAG       </head>
TAG       <p id="Omk.1.1.1">
TOK       Беше
TOK       јасен
TOK       и
TOK       студен
TOK       априлски
TOK       ден
PUN       ,
TOK       а
TOK       часовниците
TOK       отчукуваа
TOK       тринаесет
PUN_TERM          .
```

Figure 1: *The tokenized intermediate version of the text.*

Corpus encoding and structure

Taking the tokenized version of the text as a starting point, the corpus was encoded in accordance with the XML-based recommendations of the Text Encoding Initiative, TEI P4 [6]. As in MULTEXT-East, we used the *TEI.prose* base tag set and the following additional tag sets: *TEI.corpus*, which gives us the root element of the corpus and a more detailed structure of the corpus header; *TEI.linking* for pointer mechanisms; TEI.analysis for basic linguistic analysis; and *TEI.fs* for feature structures, which encode our morphosyntactic descriptions and specifications.

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE text SYSTEM "tei2.dtd">
<text id="mteo-mk." lang="mk">
 <body id="Omk" lang="mk">
  <div type="part" n="1" id="Omk.1">
   <head>I Дел </head>
    <div type="chapter" n="1" id="Omk.1.1">
     <head>I </head>
      <p id="Omk.1.1.1">
       <s id="Omk.1.1.1.1">
        <w ana="V">Беше</w>
        <w ana="A">јасен</w>
        <w lemma="и" ana="Ccs">и</w>
        <w ana="A">студен</w>
        <w ana="A">априлски</w>
        <c>,</c>
        <w lemma="ден" ana="Ncmsnn">ден</w>
        <w lemma="а" ana="Ccs">а</w>
        <w lemma="часовник"
                 ana="Ncmpny">часовниците</w>
        <w ana="M">тринаесет</w>
        <c>.</c>
        <w ana="V">отчукуваа</w>
       </s>
```

Figure 2: *The TEI structure of the novel.*

The novel is composed of three parts and an Appendix, and each of these consists of a number of chapters, marked up using the *<div>* element with the appropriate type attribute. The divisions are then composed of paragraphs (tag *<p>*), and these of sentences (tag *<s>*). All elements, down to the sentence level are given identifiers. Finally, the sentences contain words (tag *<w>*) and punctuation marks (tag *<c>*), which can be qualified by their type and linguistic annotation. This structure is shown in Figure 2.

The size of the Macedonian “1984” corpus is 3.6 MB, and is on par with the MULTEXT-East releases in other languages. The number of different tags used in the final document is set forth in Table 1.

tag	count
<div>	28
<p>	1287
<s>	6821
<c>	17075
<w>	95954

Table 1: *Tag usage of the corpus.*

Linguistic annotation

For linguistic annotation, the default *TEI.analysis* attributes *lemma* and *ana* are added afterwards from lexical lists containing the word-form, the lemma, and the MSD. For example the Macedonian wordform ден might appear in the corpus as:

```
<w lemma="ден" ana="Ncmsnn">ден</w>
```

At the moment, the first available usage of the word is taken, which effectively removes all ambiguity. Therefore, ambiguous words still have to be corrected by hand in the corpus. Also, only several part-of-speech categories have *lemma* attributes in the encoded corpus, due to the unavailability of wordlists for the other categories. Most notably, verbs and adjectives are represented with an *ana* attribute containing only a part-of-speech category, and have no *lemma* attribute.

4 ALIGNMENT

The Macedonian translation of 1984 was automatically sentence aligned with the MULTEXT-East English original and the alignment hand validated.

The aligning was done using the Vanilla aligner [7]. It is a language independent aligner and uses an algorithm which assumes that the source and its translation consist of an equal number of smaller parallel units, delimited in some known way. All it has to do is to align smaller units inside these parallel units. In our case, the paragraphs were aligned to start with; therefore the alignment problem was driven down to aligning the sentences in each paragraph. So, we assume that the number of paragraphs is the same in both texts and the paragraphs are pair wise parallel. The algorithm also assumes that the order of sentences in the original text is the same as in the translation.

The algorithm works on the basis of the assumption that the length of the original and its translation are correlated. The translations of longer sentences are longer than translations of shorter sentences. When aligning the units, one should try to achieve that the length of the original is not too different from the length of the translation. Therefore it is sometimes necessary to prefer 0-1, 1-2, 2-1 or other complicated alignments to 1-1 alignment. It is also necessary to convert the text to a fixed length encoding, as the algorithm doesn't work with variable length character encodings, such as the UTF-8 we are using. Therefore, the corpus was transliterated to a Latin script, thus enabling the proper usage of the aligner.

The result from the aligner was hand checked several times in order to ensure the correctness of the alignment. However, there were cases where the aligner output wrong alignments, which were documented, and afterwards corrected manually.

The alignments are encoded in a separate document containing references to sentence IDs, as specified by the *cesAlign DTD*, an application of the Corpus Encoding Standard [8]. Figure 3 gives a Macedonian-English alignment span illustrating the syntax and types the alignment links: the first link encodes a 1-1 alignment, the second a 1-0, and the third an 2-1 alignment.

```
<link xtargets="Omk.3.1.76.1 ; Oen.3.1.76.1"/>
<link xtargets="Omk.3.1.77.1 ;"/>
<link xtargets="Omk.3.1.77.2 Omk.3.1.77.3;
                Oen.3.1.77.1"/>
```

Figure 3: *Example of bilingual alignment.*

5 LEARNING POS TAGGING USING TnT

TnT [9], the short form of *Trigrams'n'Tags*, is a very efficient statistical part-of-speech tagger that is trainable on different languages and virtually any tagset. The component for parameter generation trains on tagged corpora. The system incorporates several methods of smoothing and of handling unknown words. TnT is optimized for training on a large variety of corpora. Adapting the tagger to a new language, or a new tagset is very easy. Additionally, TnT is optimized for speed.

The tagger is an implementation of the Viterbi algorithm for second order Markov models. The main paradigm used for smoothing is linear interpolation, and the respective weights are determined by deleted interpolation. Unknown words are handled by a suffix trie and successive abstraction. TnT was chosen primarily for its performance/speed ratio. It currently stands as one of the fastest state-of-the-art PoS taggers available.

Learning the Tagger

The tagset used was the same one we worked with during the annotation. Currently, only a few PoS categories (nouns, conjunctions, particles, and adpositions) utilize the full MSD notation, whereas all the other categories are simply tagged with the PoS category letter, making no distinction on inter-part-of-speech attributes. This has probably led to an increase in the accuracy results, but as recent studies show [10], one not so drastic as to make a huge impact on our results.

For our dataset we took the newly created Macedonian “1984” corpus. As outlined in the previous chapters, the corpus was segmented and tokenized, and each word annotated with its MSD. This fully annotated corpus was converted to a format acceptable by the TnT tagger.

This tagged corpus is then used for learning the tagger, which generates the appropriate n-gram and lexicon files. It is these that are the afterwards used for tagging unknown texts. In absence of other pre-tagged Macedonian corpora, we performed cross-validation on the “1984” corpus, in order to obtain tagging accuracy results.

Evaluation

Average state-of-the-art PoS tagging accuracy is between 96% and 97%, depending on language and tagset. Our system achieved an at least en-par accuracy.

Table 3 presents the accuracy results using 10-fold cross-validation. The results were averaged over 10 test runs, and the training and test set were disjoint and randomly picked. The table shows the percentage of unknown tokens, separate accuracies and standard deviations for known and unknown tokens, as well as the overall accuracy. Let us note the achieved accuracy of 100% for known tokens. This is due to the absence of ambiguity between PoS categories in the corpus, as it has not yet been hand tagged with the correct morphosyntactic annotations. Introduction of the correct annotations in the corpus will result with a slight decrease in the accuracies achieved.

percentage	known		unknown		overall	
unknowns	acc.	σ	acc.	σ	acc.	σ
10,99%	100%	0	83,2%	0,71	98.1%	0,22

Table 3: *Part of speech tagging accuracy.*

Figure 4 shows the learning curve of the tagger, i.e., the accuracy depending on the training data. The bottom axis shows the training set size i.e. the number of tokens used for training. Each training set size was tested ten times, the training and test sets were disjoint and picked randomly, and the results were averaged. The training length is given on a logarithmic scale.

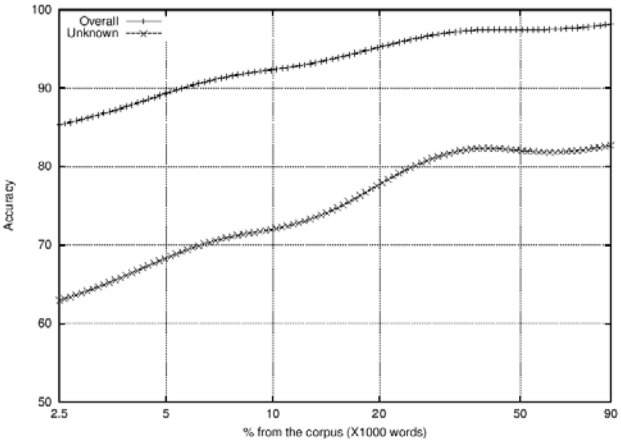

Figure 4*: Learning curve for tagging the corpus.*

6 CONCLUSION AND FURTHER WORK

In this paper we presented the process of digitalization of Orwell's “1984”, and the subsequent conversion to a standard TEI format, resulting with the first annotated corpus in Macedonian.

In addition, we described the creation of an alignment between the sentences in the Macedonian and the English editions of “1984”, therefore producing a substantial equivalence of the documents present in MULTEXT-East for the other languages.

We used our newly created resources for learning a PoS tagger for the Macedonian language. The outcome of the evaluation, albeit over incomplete data, showed promising results on par with state-of-the-art PoS tagging accuracies.

Considering further work on the subject, it is our opinion that finalization of the lexical lists, and subsequent automatic and hand tagging of the corpus is of prime importance. Re-learning the tagger with those resources would yield the real tagger performance.

It would also be of interest to see how the tagger would perform on a non-”1984” text, which illustrates the importance of creation of new annotated corpuses.

Overall, the resources created and the results obtained provide a milestone that should be built upon, and we hope that it will serve as a reference point for all kinds of Macedonian language engineering applications.

Acknowledgements

The work presented was supported by the bilateral Macedonian-Slovene project “Gathering, annotation and analysis of Macedonian-Slovene language resources”.

References

[1] T. Erjavec. *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*. In Fourth International Conference on Language Resources and Evaluation, LREC-04, Paris, 2004. ELRA. http://nl.ijs.si/et/Bib/LREC04/.

[2] K. Zdravkova et al. *Machine learning of Macedonian Nouns*. In proceedings of SIKDD 2005, Ljubljana, 2005 (to be published)

[3] Орвел. Џ, “*1984*”. Детска Радост. Skopje, 1998.

[4] Petar Kajevski. *Сто илјади најчесто користени македонски зборови.* Retrieved on 4th August from the WWW: http://www.najdi.org.mk/rang-001.html

[5] T. Erjavec et al. *Massive multilingual corpus compilation: Acquis Communautaire and totale*. In 2nd Language and Technology Conference L&T'05, Poznan, Poland, 2005.

[6] Sperberg-McQueen, C. M. and Burnard, L. (eds.). *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Bergen, 2002.

[7] Gale, W. A. and Church, K. W.. *Program for aligning sentences in bilingual corpora*. Computational Linguistics 19, pp. 75-102, 1993.

[8] Nancy Ide. *Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora*. In First International Conference on Language Resources and Evaluation, LREC'98, Granada, ELRA, 1998

[9] T. Brants. *TnT - a statistical part-of-speech tagger*. In Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, 2000.

[10] S. Džeroski, T. Erjavec, and J. Zavrel. *Morpho-syntactic Tagging of Slovene: Evaluating PoS Taggers and Tagset*s. In Second Intl. Conf. On Language Resources and Evaluation, LREC'00, Paris, 2000.