ÉTUDES COGNITIVES, 8
SOW, Warszawa 2008

LUDMILA DIMITROVA
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia

BULGARIAN DIGITAL RESOURCES AS A BASE FOR AUTOMATIC DISAMBIGUATION

Abstract

The paper describes briefly the first-ever annotated Bulgarian digital lexical resources, which were developed in the frame of EC project MULTEXT-East, and some results of an experiment in automatic part-of-speech disambiguation, based on these resources.

Introduction

Multilingual free-access language resources for research purposes have been produced in the frame of the EC projects MULTEXT, MULTEXT-East, and CONCEDE. The MULTEXT project created the first annotated large-scale multilingual corpus for seven Western European languages: Dutch, English, French, German, Italian, Spanish, and Swedish. MULTEXT-East (MTE) project is an extension of the MULTEXT project (Ide and Véronis 1994). The MULTEXT's methodology and lingualware were used for the development of language resources in six Central and Eastern European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene (Dimitrova et al. 1998). The MTE project also adapted existing tools and standards to these languages. Both projects, MULTEXT and MTE, created the first annotated large-scale multilingual corpus for 13 European languages. The MTE project built a new annotated multilingual corpus (MTE corpus), composed of material comparable to MULTEXT's. In this way, examples were provided for the applicability of, firstly, MULTEXT's multilingual tools (especially engine-based tools, alignment software, and multilingual extraction tools) to CEE languages, and secondly, the Text Encoding Initiative

(TEI) *Guidelines* and MULTEXT's TEI-based corpus markup standard to CEE languages, as well as the MULTEXT-EAGLES pan-European lexical specifications and part-of-speech (POS) tagset (Ide and Sperberg-McQueen 1995, Sperberg-McQueen and Burnard 1994).

Bulgarian MTE digital resources

The MTE digital lexical resources contain multilingual MTE corpus and a dataset of language-specific resources (Dimitrova 1998, Dimitrova et al. 2005). The first-ever annotated Bulgarian MTE digital lexical resources, which were developed in the frame of EC project MULTEXT-East, were used for various purposes and applications to language engineering. One of the significant applications was an experiment for automatic part-of-speech disambiguation. The experiment was carried out through word-level morphosyntactic markup Bulgarian translation of George Orwell's novel "1984". As a result, a text was obtained, such that every word-form in it was annotated with the most relevant POS tag. The text is thus presented in the form of a most linguistically informative document.

The MTE language-specific resources contain morphosyntactic specifications (coded as MorphoSyntactic Descriptions — MSDs) for the six CEE languages, as well as for English, and data for use with the various annotation tools, namely:

- Segmentation rules. These include rules describing the form of sentence boundaries, quotations, numbers, punctuation, capitalization, etc.
- Special tokens. The language-specific data required by the segmentation tools includes lists of special tokens (frequent abbreviations and names, titles, patterns for proper names, etc.) with their types.
- Morphological rules. The project is providing morphological rules for the MULTEXT-East languages, needed by the morphological tools. The rules provide exhaustive treatment of inflection and minimal derivation. Each lemma in the lexical lists used by the project is associated with its part(s) of speech and morphological rules.
- Lexicon (Lexical lists). For the purposes of the corpus morpho-lexical processing, the MULTEXT-East consortium developed language-specific word-form lexical lists covering at least the words appearing in the corpus. For each of the six MTE languages, as well as for English, a lexical list containing at least 15,000 lemmas is being developed, for use with the morphological analyser. Each lexicon entry includes information about the: inflected-form, lemma, POS, and morphological specifications. A mapping from the morphosyntactic information contained in

the lexicon to a set of corpus tags (used by the part-of-speech disambiguator) is also provided, according to the MULTEXT tagging model.

A lexicon entry has the following structure:

word-form <TAB> **lemma** <TAB> **MSD** <TAB> **comments**

where word-form represents an inflected form of the lemma, characterised by a combination of feature values encoded by **MSD**-code; the forth column, comments, which is optional, is currently ignored and may contain either comments or information processable by other tools.

A Bulgarian Lexicon excerpt follows:

Word-Form	**Lemma**	**MSD**
катереше	катеря	Vmii2s
катереше	катеря	Vmii3s
катери	катер	Ncmp-n
катери	катеря	Vmia2s
катери	катеря	Vmia3s
катери	катеря	Vmip3s
катери	катеря	Vmm-2s
катерите	катер	Ncmp-y
катерите	катеря	Vmip2p
катеричката	катеричка	Ncfs-y
катинарче	=	Ncns-n
като	=	Css
като	=	Sp

The morphosyntactic descriptions are provided as strings, using a linear encoding; a relatively efficient and compact way to represent the flat attribute-value matrices. 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 the one-character code; if an attribute does not apply, the corresponding position in the string contains the special marker “-“ (hyphen). By convention, trailing hyphens are not included in the lexical MSDs. Such specifications provide a simple and relatively compact encoding, and are in intention similar to feature-structure encoding used in unification-based grammar formalisms. When the word-form is the very lemma, then the equal sign is written in the lemma-field of the entry ("="). Here we can mention that the number of attributes, for example, for a POS **noun** are 10, with values for these attributes being 54; correspondingly, for a POS **verb** there are 13 attributes with 53 values; for a POS **pronoun** there are 16 attributes with 83 values; for a POS **adjective** — 12 and 61; for a POS **adverb** — 5 and 20; for a POS **numeral** — 12 and 72; for a POS **conjunction** — 7 and 21, which can be noted in the Appendix.

The MTE corpus is composed of three major parts:

(1) Multilingual Comparable Corpus

For each of the six CEE languages, the comparable corpus included two subsets of at least 100,000 words each, consisting of

- fiction, comprising a single novel or excerpts from several novels;
- newspapers.

The data was comparable across the six languages, in terms of the number and text size. The entire multilingual comparable corpus was prepared in Ces format (**Ces**: **C**orpus **E**ncoding **S**tandard), manually or using ad-hoc tools, and was automatically annotated for tokenization, sentence boundaries, and part-of-speech annotation using the project tools.

(2) Multilingual Speech Corpus

MTE records a small corpus of spoken texts in each of the six languages.

(3) Multilingual Parallel Corpus

The parallel MULTEXT-East corpus consists of six integral translations of George Orwell's "1984": besides the original English version, the corpus contains translations in Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene, and includes approximately 100,000 words per language. For each language, the corpus was marked and validated for paragraph and sentence boundaries.

Structure of MTE parallel corpus

There are four versions of this parallel corpus, corresponding to the different levels of annotation:

- the original texts,
- the CesDOC encoding,
- the CesAlign aligned versions,
- the CesANA encoding.

The ***original texts*** corpus consists of seven texts of George Orwell's *Nineteen Eighty-Four*: besides the original English text, the corpus contains translations in the six project languages.

The ***CesDOC encoding version*** contains original texts, marked-up in SGML format, (SGML: **S**tandard **G**eneralized **M**arkup **L**anguage) (Burnard 1995). SGML-markup has made for gross document structure (major text divisions), down to the level of the paragraph/sub-paragraph and the sentence boundaries: <div>-</div>, <p>-</p>, and <s>-</s> tags.

An example of the CesDOC encoding follows:

<text>

<body lang=bg id=Obg.1984>

<div id="Obg.1" type=part n=1>

<div id="Obg.1.1" type=chapter n=1>

<p id="Obg.1.1.1">

<s id="Obg.1.1.1.1"> Априлският ден бе ясен и студен, часовниците биеха тринайсет часа. </s>

<s id="Obg.1.1.1.2"> С глава, сгушена между раменете, за да се скрие от лютия вятър, <name type=person> Уинстън Смит </name> се шмугна бързо през остъклените врати на жилищен дом <name type=place rend=dblq>

Победа</наме>, но не толкова бързо, че да попречи на вихрушката прахоляк да нахлуе с него. </s></p>

<p id="Obg.1.1.2">

<s id="Obg.1.1.2.1"> В коридора миришеше на варено зеле и стари парцалени изтривалки. </s>

<s id="Obg.1.1.2.2"> На стената в единия му край бе закачен с кабърчета цветен плакат, прекалено голям за каквото и да е помещение. </s>

The ***CesAlign aligned version*** **-** the CesAlign version is associated with each of the non-English texts, which includes links between sentences (in the cesDoc encoding) for each text (non-English and English), thus providing a parallel alignment at the sentence level. For each language, the corpus was marked and validated for alignment. Alignment between each of the six CEE languages and the English text ensures six pair-wise alignments. The alignments were obtained by two different aligners (MULTEXT-aligner and Vanilla aligner), with accuracy ranging between 75 and 80% and were afterwards hand-validated and corrected. For Bulgarian, the alignment was made by the Vanilla aligner. The table below shows the distribution of Bulgarian-English sentences alignment, 6699 bilingual links in total:

Aligned Bulgarian-English pairs	**Nr.**	**Proc**
2-2	2	0.030017%
2-1	23	0.345190%
1-2	36	0.540297%
1-1	6637	99.074487%
0-1	1	0.014970%

Example. The MTE aligned version for Bulgarian *(Bulgarian-English aligned "1984")* 1-1 aligned correspondence sampler follows:

- -

<**Obg.**1.1.1.1>Априлският ден бе ясен и студен, часовниците биеха тринайсет часа.
<**Oen.**1.1.1.1>It was a bright cold day in April, and the clocks were striking thirteen.

- -

<**Obg.**1.1.1.2>С глава, сгушена между раменете, за да се скрие от лютия вятър, **Уинстън Смит** се шмугна бързо през остъклените врати на жилищен дом **Победа**, но не толкова бързо, че да попречи на вихрушката прахоляк да нахлуе с него.
<**Oen.**1.1.1.2> **Winston Smith**, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of **Victory Mansions**, though not quickly enough to prevent a swirl of gritty dust from entering along with him.

- -

The ***CesANA encoding version*** — the Orwell corpus is available as a tokenised and morphosyntactically tagged cesAna document. The list of cesAna encoding elements for tokens and morphosyntactic annotation encoding includes:

- <tok> contains a token, consisting of its orthographic form in the original document, followed optionally by disambiguated corpus tag and one or more options of morphosyntactic information associated with the token, tokens are of TYPE=WORD or PUNCT;
- <orth> contains the orthographic form of the token as it appears in the original;
- <disamb> contains one disambiguated corpus tag associated with the token, i.e. context-dependent, disambiguated lexical information;
- <lex> contains one or more options of morphosyntactic information associated with the token — ambiguous lexical information;
- <base> the base or lemma of a token (lemmatized form for the morphosyntactic information given in the associated <msd> element);
- <msd> the MSD of a token;
- <ctag> contains the corpus tag associated with the morphosyntactic information.

How to obtain a cesana encoded document

To arrive at the tokenised and tagged *cesAna Orwell document* (for example, a Bulgarian Orwell's "1984") the following steps have been performed:
1. cesDoc version has been simplified and converted to cesAna encoding,
2. the text (the result of step 1.) was tokenized,
3. the tokens (the result of step 2.) were annotated with lexical (ambiguous MSDs) lemmas and tags,
4. the lexical information was disambiguated.

The software tools with which the above-mentioned steps were carried out were developed within the MULTEXT project, but the data input came from MTE language-specific resources.

At first, the Bulgarian text from the MTE corpus was segmented by means of the segmenter MTSeg — a tokenizer. The segmenter is a language-independent and configurable processor used to tokenize input text, given in one of the three possible formats: plain text, a normalized SGML form (nSGML) as output by another MULTEXT tool (MTSgmlQl), or a tabular format (also specific to MULTEXT processing chain). The output of the segmenter is a tokenized form of the input text, with paragraph and sentence boundaries marked-up. Punctuation, lexical items, numbers and several alphanumeric sequences (such as dates and hours) are annotated with various tags out of a hierarchy class structured tag set. The language specific behavior of the segmenter is driven by several language resources (abbreviations, compounds, split words, etc.), incl. segmentation rules and special tokens.

To explain the structure of the final documents, first consider a fragment of the English cesDoc Orwell:

```
<p id="Obg.1.1.1">
<s id="Obg.1.1.1.1">
        Априлският ден бе ясен и студен, часовниците биеха тринайсет часа.
</s>
<s id="Oen.1.1.1.2">
С глава, сгушена между раменете, за да се скрие от лютия вятър,
    <name type=person>Уинстън Смит</name>,
се шмугна бързо през остъклените врати на жилищен дом
<name type=place>Победа</name>,
        но не толкова бързо, че да попречи на вихрушката прахоляк да нахлуе с него.
</s>
</p>
```

At the S (**S***entence*) level the documents have been tokenised according the lexical resources of the language and are encoded as TOKen elements. Tokens are either "normal" words, compounds, separable parts of words ("*clitics*"), or punctuation marks. They are distinguished by the value of the token's TYPE attribute. The values used are WORD for words, and PUNCT for punctuation marks. The word or punctuation mark is contained in the ORTH element. The punctuation tokens are annotated with (unambiguous) corpus tags, which identical across the languages of MULTEXT-East.

The following example illustrates this markup:

```
<par from=”Obg.1.1.1”>
<s from=”Obg.1.1.1.1”>
<tok type=WORD><orth> Априлският </orth></tok>
<tok type=WORD><orth>ден</orth></tok>
<tok type=WORD><orth>бе</orth></tok>
<tok type=WORD><orth>ясен</orth></tok>
<tok type=WORD><orth>и</orth></tok>
<tok type=WORD><orth>студен</orth></tok>
<tok type=PUNCT><orth>,</orth><ctag>COMMA</ctag></tok>
<tok type=WORD><orth> часовниците </orth></tok>
<tok type=WORD><orth>биеха</orth></tok>
<tok type=WORD><orth> тринайсет </orth></tok>
<tok type=WORD><orth>часа</orth></tok>
<tok type=PUNCT><orth>.</orth><ctag>PERIOD</ctag></ctag>
</tok>
</s>
```

When the input text was segmented, the next tool — MTLex — from MULTEXT tools was used: a dictionary look-up procedure assigns to each lexical token all its possible *morpho-syntactic descriptors* (MSDs). Corresponding lines for morphosyntactic annotation of the Bulgarian phrase *“ден бе”* in output of MTLex (in English “*day was*” — from the first sentence of the “1984”: *It **was** a bright cold **day** in April, and the clocks were striking thirteen.*) are:

```
1.2.1.1.1.1.1.1\12TOK   ден   ден\Ncms-\NCMS-N
1.2.1.1.1.1.1.1\16TOK    бе   бе\Qgs\QGS|съм\Vaia2s\VAIA2S|съм\Vaia3s\VAIA3S
```

At the next step the text was tokenized. The word tokens are annotated both with ambiguous lexical information (in the <lex> elements of the token), and with disambiguated, context-dependent, information (in the <disamb> element(s)). Both elements contain the <base> (lemma) of the token, its morphosyntactic description <msd>, and its language depended corpus tag — <ctag> — as illustrated in the following example, the first sentence of the Bulgarian translation of “1984” — *Априлският ден бе ясен и студен, часовниците биеха тринайсет часа.* (In English: *It **was** a bright cold **day** in April, and the clocks were striking thirteen.*)

```
<par from=‘Obg.1.1.1’>
  <s from=‘Obg.1.1.1.1’>
    <tok type=WORD from=‘Obg.1.1.1.1\1’>
      <orth Априлският </orth>
```

```
<disamb><base>априлски </base><ctag>AMS</ctag></disamb>
      <lex><base> априлски</base><msd>A–ms-f</msd><ctag>AMS</ctag>
                                                                    </lex>
    </tok>
    <tok type=WORD from=‘Obg.1.1.1.1\12’>
      <orth> ден </orth>
      <disamb><base> ден </base><ctag>NCMS-N</ctag></disamb>
      <lex><base> ден </base><msd>Ncms-n</msd><ctag>NCMS-N</ctag>
                                                                    </lex>
    </tok>
    <tok type=WORD from=‘Obg.1.1.1.1\16’>
      <orth> бе </orth>
      <disamb><base>съм </base><ctag>VAIA3S</ctag></disamb>
      <lex><base> бе </base><msd>Qgs</msd><ctag>QG</ctag></lex>
      <lex><base> съм </base><msd>Vaia2s</msd><ctag>VAIA2S</ctag>
                                                                    </lex>
      <lex><base> съм </base><msd>Vaia3s</msd><ctag>VAIA3S</ctag>
                                                                    </lex>
    </tok>
    <tok type=WORD from=‘Obg.1.1.1.1\19’>
      <orth> ясен </orth>
      <disamb><base> ясен </base><ctag>AMS</ctag></disamb>
      <lex><base> ясен </base><msd>A–ms-n</msd><ctag>AMS</ctag>
                                                                    </lex>
      <lex><base> ясен </base><msd>Ncms-n</msd><ctag>NCMS-N</ctag>
                                                                    </lex>
    </tok>
    <tok type=WORD from=‘Obg.1.1.1.1\24’>
      <orth> и </orth>
      <disamb><base> и </base><ctag>CC</ctag></disamb>
      <lex><base> и </base><msd>Ccs</msd><ctag>CC</ctag></lex>
      <lex><base> и </base><msd>I-s</msd><ctag>I</ctag></lex>
    </tok>
    <tok type=WORD from=‘Obg.1.1.1.1\26’>
      <orth> студен </orth>
      <disamb><base> студен </base><ctag>AMS</ctag></disamb>
      <lex><base> студен </base><msd>A–ms-n</msd><ctag>AMS</ctag>
                                                                    </lex>
    </tok>
    <tok type=PUNCT from=‘Obg.1.1.1.1\32’>
      <orth>,</orth>
      <ctag>COMMA</ctag>
    </tok>
    <tok type=WORD from=‘Obg.1.1.1.1\34’>
      <orth>часовниците</orth>
      <disamb><base>часовник</base><ctag>NCMP-Y</ctag></disamb>
      <lex><base>часовник</base><msd>Ncmp-y</msd><ctag>NCMP-Y<
                                                             /ctag></lex>
    </tok>
    <tok type=WORD from=‘Obg.1.1.1.1\46’>
      <orth>биеха</orth>
      <disamb><base>бия</base><ctag>VMII3P</ctag></disamb>
      <lex><base>бия</base><msd>Vmii3p</msd>
          <ctag>VMII3P</ctag></lex>
    </tok>
    <tok type=WORD from=‘Obg.1.1.1.1\52’>
      <orth>тринайсет</orth>
      <disamb><base> тринайсет </base><ctag>MC</ctag></disamb>
      <lex><base> тринайсет </base><msd>Mc-p-ln</msd><ctag>MC</ctag>
                                                                  </lex>
    </tok>
    <tok type=WORD from=‘Obg.1.1.1.1\63’>
      <orth>часа</orth>
      <disamb><base>час</base><ctag>NCMS-S</ctag></disamb>
      <lex><base>час</base><msd>Ncms-s</msd><ctag>NCMS-S</ctag>
                                                                  </lex>
      <lex><base>час</base><msd>Ncmt</msd><ctag>NCMT</ctag></lex>
    </tok>
      <tok type=PUNCT from=‘Obg.1.1.1.1\67’>
      <orth>.</orth>
      <ctag>PERIOD</ctag>
    </tok>
  </s>
```

Let us look at the Bulgarian phrase *“ден бе”* (in English “*day was*” — from the first sentence of the “1984”):

```
<tok type=WORD from=‘Obg.1.1.1.1\12’>
  <orth>ден</orth>
  <disamb><base>ден</base><ctag>NCMS-N</ctag></disamb>
  <lex><base>ден</base><msd>Ncms-n</msd><ctag>NCMS-N</ctag></lex>
</tok>
<tok type=WORD from=‘Obg.1.1.1.1\16’>
    <orth>бе</orth>
    <disamb><base>съм</base><ctag>VAIA3S</ctag></disamb>
    <lex><base>бе</base><msd>Qgs</msd><ctag>QG</ctag></lex>
    <lex><base>съм</base><msd>Vaia2s</msd><ctag>VAIA2S</ctag></lex>
    <lex><base>съм</base><msd>Vaia3s</msd><ctag>VAIA3S</ctag></lex>
</tok>
```

We should now state that the annotation is correct both in the <disamb>, as well as in the <lex> elements. The <lex> elements of the token represent its ambiguity class (only one of the <lex> elements is correct in the given context!). In cases where the tagger or human could not decide how to disambiguate there may be more <disamb> elements for one token. In this case each <disamb> element appears among the <lex> elements as well.

The table shows some statistical data of tag usage in Bulgarian and English Orwell's "1984":

Lang	**tokens**	**words**	**disamb**	**lex**	**base**	**MSD**
BUL	101173	86020	86020	156002	242022	156002
ENG	118102	103997	187526	214404	401930	401930

As is clear, the Bulgarian version contains: 156002 units of ambiguated lexical information; such is the number of MSDs of the tokens as well; bases or lemmata of the tokens are 242022.

For Bulgarian the relation between the number of the **words** (equal to the units of disambiguated lexical information), **lex** (equal to MSD), and **base** is: number of the **words** *plus* number of the **lex** equals the number of **base** ($86020 + 156002 = 242022$).

Automatic disambiguation of bulgarian Orwell's "1984"

Automatic disambiguation is a very important application for many natural language processing tasks. Assigning of the correct lemma (base form) to each word-form in an annotated text is not trivial for Bulgarian, as there is no statistical data that can provide disambiguation of the annotated text. Geneva's ISSCO tagger 2.22 was used for Bulgarian. ISSCO tagger 2.22 is a POS disambiguator; it takes as input the output of the morphological analyser, consisting of word-forms associated with one or several possible POS tags (with morphological information, and lemma).

The linguistic data resources for ISSCO tagger 2.22 are provided by two language specific tables: tbl.tag.corpus.bg, containing MSDs, and states.list, containing Corpus tags (Ctags). (A tag is a specific identifier — a string of characters identifying an element — that marks the boundaries of an element. A Ctag contains the corpus tag associated with the morphosyntactic information, i.e. provides POS-information in the form of a corpus tag.)

The file "tbl.tag.corpus.bg" contains 326 elements distributed in two columns: the 326 MSDs in its first column and the corresponding 326 Ctags in its second column. The correspondance is 1-1 (see bellow an excerpt for a POS Verb):

Vmpa-sfa-n	VMPA-SFA-N
Vmip3p	VMIP3P
Vmip3s	VMIP3S
Vmm-2p	VMM-2P
Vmm-2s	VMM-2S
Vmpa-p-a-n	VMPA-P-A-N
Vmpa-p-a-y	VMPA-P-A-Y

The table tbl.tag.corpus.bg was used to prepare the file "states.list" (the list of Ctags for Bulgarian) the ISSCO tagger 2.22 works with. In order to run the tagger for disambiguation of the annotated text, we had to reduce the number of Ctags. Since some attributes of the MSDs do not apply for Bulgarian, the number of the Ctags was further reduced by dropping off some positions in the MSDs. Generally, the principle of the reduction was to exclude these features that are not specific for Bulgarian, without losing information. Such features occur mostly in adjectives, pronouns and numerals. The reduction did not affect the descriptions of nouns and verbs. For example, the MSD of an adjective of type "A–ms-y" (adjective, masculine, singular, definite) was reduced to Ctag "ASM" (adjective, singular, masculine), and the MSD of an adjective of type "A–ms-n" was reduced to Ctag "AMS". We thus reduced the nine MSDs of a POS adjective to four Ctags of a POS adjective:

A–p-n	**AP**	A–ms-n	**AMS**
A–p-y	**AP**	A–ms-f	**AMS**
A–fs-y	**AFS**	A–ms-s	**AMS**
A–fs-n	**AFS**	A–ns-y	**ANS**
		A–ns-n	**ANS**

After we reduced the number of Ctags to 117 in this way, we filled out table states.list as a resource for the tagging program.

The role of the POS disambiguator is to select the most plausible POS tag on the basis of the local context. The process is based on a Markov model that selects the most plausible tag using statistical generalization given the categories of the two preceding words. The statistical method is: accurate (90-96% correct), efficient (linear time in proportion to the input), language-independent, etc. The process of disambiguation is accomplished in *two steps*: a *training phase* to estimate the parameters of the model and a *testing phase* to select the most probable tags according to this model (employing the Viterbi algorithm). The tagger was trained on a manually annotated text; in such cases the annotation of input data for training phase is unambiguous. The output of each cycle of the training phase is a matrix filled-in with data of the most relevant tag(s). To realize an automatic disambiguation of the annotated Bulgarian text by ISSCO tagger 2.22 we carried out two experiments to train the tagger. In the experiments, we used the manually tagged chapters 1, 2, and 3 of part 1 and chapter 1 of part 2 of “1984”.

First experiment

In the training process, the Matrix MM has been created by a manually tagged input text from chapter 2 of part 1, chapter 3 of part 1 (partially), and chapter 1 of part 2 with a total word count of 7300, constituting 8.53% of “1984”.

The 12 consecutive cycles of training with 700 words from chapter 3 of part 1 gave us the following error rates for disambiguation:

1 cycle: input matrix MM, 3.20% error rate, new matrix MM1
2 cycle: input matrix MM1, 3.01% error rate, new matrix MM2
. . .
6 cycle: input matrix **MM5, 2.95%** error rate, new matrix MM6
. . .
11 cycle: input matrix MM10, 3.01% error rate, new matrix MM12
12 cycle: input matrix MM11, 3.06% error rate, new matrix MM13

As we observed an increase in the error rate after the 6th cycle, we chose for testing phase MM5, which gave us a minimum error rate.

For a testing corpus we used the manually tagged full text of chapter 1 of part 1 with a total word count of 5737, constituting 6.68% of “1984”. The number of Ctags was 117, of MSDs — 326, separately from the set of punctuation tags.

The statistics data we obtained after the testing phase with the manually tagged chapter 1 of part 1 (word count 5737) and matrix MM5 is: Word count — 5737, Error count — 286. This gives us an error rate of 4.99 %.

(The testing phase with the same manually tagged text and matrix MM10 gives us an error count of 291 for word count of 5737, and an error rate of 5.07 %.)

Second experiment

In the training process, the Matrix MM has been created by the manually tagged input text from chapter 1 of part 1, chapter 2 of part 1 (partially), and chapter 1 of part 2 with a total word count of 12338, constituting 14.21% of “1984”.

The 9 consecutive cycles of training with 3237 words from chapter 3 of part 1 gave us the following error rates for disambiguation:

1 cycle: input matrix MM, 5.79% error rate, new matrix MM1
2 cycle: input matrix MM1, 5.40% error rate, new matrix MM2
3 cycle: input matrix MM2, 5.44% error rate, new matrix MM2
4 cycle: input matrix **MM3, 5.33%** error rate, new matrix MM2
5 cycle: input matrix MM4, 5.37% error rate, new matrix MM2
6 cycle: input matrix MM5, 5.47% error rate, new matrix MM6
7 cycle: input matrix MM6, 5.58% error rate, new matrix MM2
8 cycle: input matrix MM7, 5.61% error rate, new matrix MM2
9 cycle: input matrix MM8, 5.61% error rate, new matrix MM2

As we observed an increase in the error rate after the 4th cycle, the best matrix from the second experiment is MM3. It gave us a minimum error rate.

Its error rate of 5.33% is significantly higher than 2.95%! Therefore, we used the best matrix MM5 from the first experiment to obtain the disambiguated version of "1984".

The ambiguously MSD-annotated texts and their corresponding disambiguated ones were the basis for building the CesANA encoded version of the multilingual parallel corpus. Some statistics from the Bulgarian CesANA document follows: annotated corpus Orwell's "1984" consists of 87235 words, where distinct words (lemmata) are 15041, distinct MSDs in the text are 324, and distinct Ctags — 117. Let us look again at the Bulgarian phrase *"ден бе"* in Obg.CesAna-format (In English "*day was*" — from the first sentence of the "1984": *It **was** a bright cold **day** in April, and the clocks were striking thirteen.*):

(1) Regarding ***ден***

```
<tok type=WORD from='Obg.1.1.1.1\12'>
  <orth>ден</orth>
  <disamb><base>ден</base><ctag>NCMS-N</ctag></disamb>
  <lex>
    <base>ден</base>
    <msd>Ncms-n</msd>
    <ctag>NCMS-N</ctag>
  </lex>
</tok>
```

We could summarize that for the noun *ден* (given by one disambiguation with one base, and one lex with one base and one MSD, i.e. the 1-1-mapping) the tagger gives us an unambiguous decision.

(2) Regarding ***бе***

```
<tok type=WORD from='Obg.1.1.1.1\16'>
    <orth>бе</orth>
    <disamb><base>съм</base><ctag>VAIA3S</ctag></disamb>
    <lex><base>бе</base><msd>Qgs</msd><ctag>QG</ctag>
    </lex>
    <lex><base>съм</base><msd>Vaia2s</msd>
    <ctag>VAIA2S</ctag></lex>
    <lex><base>съм</base><msd>Vaia3s</msd>
    <ctag>VAIA3S</ctag></lex>
</tok>
```

We could summarize that for the verb-form *бе* (given by one disambiguation with one base, and three lex with three bases and three different MSD, i.e. the mapping is not 1-1) the tagger gives us an ambiguous decision.

The resulted tagging accuracy for Bulgarian was close to 95.01%, for the entire text of the Bulgarian translation of "1984" (it is normal that human taggers disagree up to 3.5%).

Conclusion

This article briefly reviews experiments for automatic part-of-speech disambiguation, based on Bulgarian language resources, carried out through an annotated Bulgarian text. These language resources were developed for the first time in Bulgarian in the framework of multilingual research projects of the European Commission MULTEXT-East. The projects have succeeded in providing foundational resources for work in Language Engineering in Bulgarian, for morphological, grammatical, semantic or other research, or as the basis for development of new applications in natural language processing.

Acknowledgements

I would like to thank all colleagues with whom I worked throughout the years for the development of the Bulgarian annotated language resources: Lydia Sinapova, Radoslav Pavlov, and Kiril Simov (Bulgarian Academy of Sciences, Sofia, Bulgaria), Tomaž Erjavec (JSI, Ljubljana, Slovenia), Vladimir Petkevič (Charles University, Prague, Czech Republic), Dan Tufiş (RACAI, Romanian Academy, Bucharest, Romania), Heiki-Jan Kaalep (University of Tartu, Tartu, Estonia), Csaba Oravecz and Laszlo Tihanyi, (Hungarian Academy of Sciences, Budapest, Hungary). I would like to direct a special thanks to the coordinator of the project Jean Véronis (CNRS-France), and to Nancy Ide (Vassar College, USA) — for the guidance and the expert help.

References

Burnard L. (1995) What is SGML and How Does It Help? *In Computers and the Humanities*, 29, pp. 41–50.

Dimitrova L. (1998) Lexical Resource Standards and Bulgarian Language. *In International Journal Information Theories & Applications.* Vol. 5, Nr. 1. Pages 27–34.

Dimitrova L., Erjavec T., Ide N., Kaalep H.-I., Petkevic V, Tufis D. (1998) Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. *Proceedings of COLING-ACL '98.* Montréal, Québec, Canada, pp. 315–319.

Dimitrova L., Pavlov, R, Simov, K, Sinapova L. (2005) Bulgarian MULTEXT-East Corpus — Structure and Content. *In Cybernetics and Information Technologies.* Vol. 5. Nr. 1, pp. 67–73.

Ide N., Sperberg-McQueen C. M. (1995) The TEI: History, Goals, and Feature. *In Computers and the Humanities*, 29, pp. 5–15.

Ide N., Véronis J. (1994) Nancy Ide and Jean Véronis. Multext (multilingual tools and corpora). *In Proceedings of the 15th COLING.* Kyoto, 1994, pp. 90–96.

Sperberg-McQueen C. M., Burnard L. ed. (1994) *Guidelines for Electronic Text Encoding and Interchange.* Chicago and Oxford.

ISSCO tagger
`http://www.issco.unige.ch/staff/robert/tatoo/tagger.html#design`

Appendix

Conjunction (C)

8 Positions, 7 attributes, 21 values

=	=======	========	=	EN	RO	SL	CS	BG	ET	HU
P	ATT	VAL	C	x	x	x	x	x	x	x
=	=======	========	=							
1	Type	coordinating	c	x	x	x	x	x	x	x
		subordinating	s	x	x	x	x	x	x	x
		portmanteau	r		x					
********************************										
2	Formation	simple	s		x	x		x		x
		compound	c		x	x		x		x
-	-----------	-------------	-							
3	Coord_Type	l.s simple	s		x					
		l.s. repetit	r		x					
		l.s. correlat	c		x					
		l.s. sentence	p							x
		l.s. words	w							x
		l.s. initial	i	x						
		l.s. non_initial	n	x						
-	-----------	-------------	-							
4	Sub_Type	negative	z		x					
		positive	p		x					
-	-----------	-------------	-							
5	Clitic	no	n		x					
		yes	y		x					
-	-----------	-------------	-							
6	Number	singular	s				x			
		plural	p				x			
-	-----------	-------------	-							
7	Person	first	1				x			
		second	2				x			
		third	3				x			
=	=======	========	=	EN	RO	SL	CS	BG	ET	HU