BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES · Volume **5**, No 1

Sofia. 2005

Bulgarian MULTEXT-East Corpus – Structure and Content

Ludmila Dimitrova, Radoslav Pavlov*, Kiril Simov**, Lydia Sinapova****

** Institute of Mathematics and Informatics, 1113 Sofia*
*** Institute for Parallel Processing, 1113 Sofia*
**** Institute of Information Technologies, 1113 Sofia; Simpson College, IA, USA*

Abstract. *The first Bulgarian language electronic corpus is included in the MULTEXT-East (MTE) multilingual corpus of the MTE-project COP 106. The Bulgarian corpus is developed in the framework of MTE according to the methodology and requirements of the project.*

Keywords: *language resources, lexical databases, morpho-syntactic descriptors, annotated multilingual corpus.*

1. Introduction

For most Western languages large-scale standardised language resources have been developed and are available for purposes of language engineering and industries or are under development, see EU LRE project MULTEXT, [2]. The MULTEXT-East (MTE) project *Multilingual Text Tools and Corpora for Central and Eastern European Languages* is a continuation of MULTEXT project under the INCO-Copernicus programme.

Project MULTEXT has produced language resources and a freely available set of tools that is extensible, coherent, and language independent, for six western European languages (English, French, Dutch, Italian, German, and Spanish). These tools have been implemented under UNIX. They could be distributed in two main types: corpus annotation tools and corpus exploitation tools – *segmenter, morphological analyzer, part-of-speech disambiguator, aligner,* etc. All tools are integrated via a common user interface into a general-purpose manipulation system suitable for natural language processing research. The MULTEXT tools have been designed with an engine-based approach where all language-dependent materials are provided as data (in a form of

the tables or rules). Therefore, the extension of the tools in MULTEXT-East to cover CEE languages is general and involves providing the appropriate tables and rules. However, some adjustments have been made to answer the problems posed by different families of languages.

Adapting existing tools and standards, MTE developed significant language resources for six Central and Eastern European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, [1]. The MTE resources – which comprise marked-up texts in six CEE languages totaling approximately 2 million words and a small speech corpus – are a valuable database for studying word-class syntactic tagging, bi-lingual lexicon extraction and other issues relevant to language engineering applications for a number of Central and Eastern European Languages.

MTE language resources include three main parts: lexicons, parallel corpus, based on George Orwell's novel "1984", and comparative corpus – newspaper excerpts and texts from CEE literatures. The texts of the parallel corpus have been produced as well-structured lemmatized documents, according to Corpus Encoding Specification (CES-corpus), [3]: each word-form is associated to the relevant grammatical information and corresponding lemma that form its standard lexical description. The texts and the lexicons produced serve as input data for experiments with the tools created for processing Western-European languages in MULTEXT**.** They also serve as resources for building lexical databases for the six CEE languages.

In this way the scope of MULTEXT has been extended to CEE languages with the goals of developing morpho-lexical resources, building an annotated multilingual corpus (parallel and comparative corpora for seven languages) tested by MULTEXT tools, and adaptation of language specifications to the international standards. The free word order and rich inflection of CEE languages (especially three Slavic: Bulgarian, Czech, and Slovene) present significantly different linguistic problems than do those of Western Europe.

The main results of the MTE project are morpholexical resources (specifications for lexicon encoding) and annotated multilingual corpus for the seven languages: CEE and English.

2. A brief description of parallel MTE corpus

In the context of MULTEXT-EAST project**,** a parallel corpus has been developed. Its main centrepiece is the George Orwell's novel "1984" in the English original and the six translations in Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene of the novel.

The text pairs Bulgarian-English, Czech-English, etc. from the novel are sentence-aligned and their words annotated for context disambiguated lemmas and morphosyntactic descriptions.

In the framework of the MTE project, a Corpus Encoding Specification, suited for use in language engineering applications, has been developed, based on the Text Encoding Initiative (TEI) group's *Guidelines for Electronic Text Encoding and Interchange*. The CES identifies a minimal encoding level that corpora must achieve in order to be considered standardized in terms of descriptive representation (marking of structural and linguistic information) that also provides encoding conventions for

linguistic annotation. The process of its application to new texts in new languages consisted in revision and extension of the CES by the relevant partner team.

The MTE parallel corpus contains four parts, corresponding to the different levels of annotation: the original text of the novel, the CesDOC-encoding (SGML mark-up of the text up to the sentence-level), the CesANA-encoding (containing word-level morpho-syntactic mark-up), and the aligned versions in CesAlign-encoding (containing links to the aligned sentences).

The table below exemplifies CesANA-encoding for Bulgarian and English. The table describes the main characteristics per language**.** The **words** column shows the number of the lexical items, excluding punctuation, the **tokens** column shows the number of words and punctuations, the **disamb** column contains disambiguated lexical information, **lex** contains the number of the undisambiguated lexical information (MSDs), **base** is base or lemma of a token, and **msd** is morpho-syntactic description of a token.

Language	words	tokens	disamb	lex	base	msd
Bulgarian	86 020	101 173	86 020	156 002	242 022	156 002
English	103 997	118 102	187 526	214 404	401 930	401 930

The entire text of the MTE corpus is encoded as a CesCorpus-element. Each of the Ces-elements comprises a Ces-header, describing the file, the source of the corpus text, the corpus encoding and its revision history.

3. Bulgarian language corpus in MTE project

3.1. A brief overview

The Bulgarian language corpus contains Bulgarian translation of Orwell's novel, Bulgarian-English aligned text, "1984"'text, tagged with morpho-syntactic descriptions, a lexicon, and newspaper excerpts and texts from contemporary Bulgarian literature. The lexical description for Bulgarian is in accordance with the terminology and the methodology used by the MULTEXT project. The electronic version of the Bulgarian translation of Orwell's novel was created manually from the printed edition, because the electronic version of the text didn't exist. As all other texts within the MTE corpora, it can be used for research/academic purposes only. The Bulgarian version of "1984" has 87 235 words. The text was spell-checked. As a part of the MULTEXT-East multilingual corpus, the Bulgarian language corpus has been prepared in a TEI-like SGML format, the Corpus Encoding Specification (CES).

The text itself is marked up for structural data (divisions, heads, footnotes, paragraphs, etc.) and, depending on the particular text, for several sub-paragraph markups e.g. abbreviations, names, quotes, highlighted material, etc. The parallel corpus is also marked up for sentences. The additional markup of the "1984" corpus, namely the alignment and tagging with morpho-syntactic descriptions has not been included in the cesDoc primary data, but is encoded in separate documents, with hyperlinks (ID references) connecting the primary data with the annotation. The alignments with English are encoded as cesAlign documents, which contain only links to the aligned sentences. The tokenised corpus components annotated with morpho-syntactic descriptions are encoded as cesAna documents. These not only have hyperlinks to the cesDoc, but also directly include tokenised and annotated primary data.

3.2. Structure of the Bulgarian “1984” corpus

The Bulgarian “1984” corpus body consists of four . The elements have the *n* attribute, giving the successive number of the appropriate level of the (except for the appendix, where n=APPENDIX). The elements have also “id” attribute, whose value has the prefix “Obg” and the chapter numbers separated by periods, e.g. <div type=chapter n=1 id=”Obg.1.2">. Counting of chapters starts from 1 in every part. Each part (except) is further subdivided into a number of . In the Bulgarian version, only the appendix is followed by a <head>.

Further the text is segmented into paragraphs (1321 -tags occur in the text), with the <quote>, <note>, and <poem> elements marked-up at the paragraph level. Sub-paragraph tagging consists of <item>, <l>, <list>, <q>, <s>, <abbr>, <date>, <foreign>, <hi>, <mentioned>, <name>, <num>, <ptr>, and <title>.

The text has been automatically sentence-segmented by MTSeg (6649 <s> tags have been inserted automatically), and the segmentation has been hand-validated.

The <q> tag is used to mark quoted dialogue. Some <q> elements have been split in this process; these are marked with type=MI, for “machine inserted”. For the <q> tag, the attribute “broken=yes” is used when no sentence-terminating punctuation appears between two dialogue fragments by the same speaker (either inside the <q> itself or in the intervening text between two <q> tags). The <name> tag is used for all proper nouns and noun phrases, denoting names. Adjectives derived from proper nouns are not tagged. All <name> tags contain the “type” attribute: “type=person”, “type=place”, etc. The tag <foreign> is used only for those Newspeak words, which are typographically distinguished in the printed version of the text, and for the Latin words. Rendering information is included within the appropriate tags where necessary as a descriptive value of the “rend” attribute. Two type of rendition are used: “rend=mdash” and “rend=dblq”. No default rendition is used. The tags <body>, , <item>, <l>, <list>, <note>, , <poem>, <q>, <quote>, <ptr> and <s> have been automatically supplied with “id” attribute with the prefix “Obg” and subsequent numbers, separated by periods, showing their hierarchical position within the text.

The entire book is marked up using the same level of detail, i.e. no part is more detailed than the rest. All tags are used in harmony with the English “1984” for MTE; the differences are due only to the differences between the English electronic text and Bulgarian printed text.

The following is an example from the Bulgarian “1984” corpus:

<text>
<body lang=bg id=bg1984>

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

<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> <s id=“Obg.1.1.2.3”> Изобразяваше само едно огромно

лице, повече от метър широко: лице на около четирийсет и пет годишен мъж с гъсти черни мустаци и сурови красиви черти.</s> <s id="Obg.1.1.2.4"><name type=person> Уинстън </name>се запъти към стълбите.</s>

3.3. The bulgarian parallel corpus

The texts from the corpora were segmented by means of the segmenter MTSeg – a tool developed within the MULTEXT project. 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) and a tabular format (specific forms to another MULTEXT tools). The output of the segmenter is a tokenized form of the input text, with paragraph and sentence boundary 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.).

After the input is segmented, a dictionary look-up procedure assigns to each lexical token all of its possible Morpho-Syntactic Descriptors (MSDs). The dictionary look-up is accomplished by MTLex tool – a part of MULTEXT tools. The ambiguous MSD-annotated texts are then disambiguated (entirely for some languages and partially for the others) by hand or by means of some different program packages.

For Bulgarian statistics that could provide disambiguation of the annotated text did not exist. That is why we used Geneva's ISSCO tagger 2.22 for Bulgarian (the same package was used for Estonian) – the automated disambiguation ensures accuracy up to 95.01% for Bulgarian language. The results were checked and corrected by hand**.**

In order to run the tagger for disambiguation of the annotated text, we had to reduce the number of Corpus tags (Ctags). Dropping off some positions in the MSDs further reduced the 326 elements from the Bulgarian MSDs. The general principle of the reduction was to exclude features not specific to the Bulgarian language without losing information.

For example, the adjective MSD "A--ms-n" was reduced to Ctag "AMS". So we reduced the nine adjective MSDs to four adjective Ctags:

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

The reduction did not affect the nouns' and verbs' MSDs. Thus 117 Ctags were obtained as a resource for the tagging program ISSCO tagger 2.22.

The tagging experiments are incomparable across the languages, because they depend very much on the programs used and on the size of the training corpus.

The ambiguously MSD-annotated texts and their corresponding disambiguated ones were the basis for building the CesANA encoded version of the multilingual parallel corpus.

3.4. Aligned corpus

Another very useful processed form of the multilingual parallel corpus is represented by the six language alignments. Alignment is between the English version and each of the six MULTEXT-EAST languages. In MTE three different aligners with accuracy

ranging between 75-80%, were used: MULTEXT-aligner, Vanilla-aligner, and Silfide-aligner. The produced links were hand-validated and corrected. For the Bulgarian-English alignment, we used the program package Vanilla-aligner. At first, the texts were aligned to the paragraph level, and then the alignments were hand-validated and corrected. At the second processing step the texts were aligned to the sentence level and were again checked and corrected by hand. The result was 6699 alignment links. The table below shows the distribution of sentence alignments for the Bulgarian-English data.

Aligned sentence	Number	Percent
1-1	6637	99.074487%
1-2	36	0.540297%
2-1	23	0.345190%
2-2	2	0.030017%
0-1	1	0.014970%

3.5. Examples

***1-1** Aligned sentence*

- **<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.

***2-1** Aligned sentence*

- **<Obg.1.1.39.1>**Това ставаше винаги през нощта. **<Obg.1.1.39.2>** Арестуваха неизменно през нощта.
- **<Oen.1.1.41.1>**It was always at night – the arrests invariably happened at night.

4. Multilingual comparable corpus

MULTEXT-EAST produced a second multilingual comparable corpus, which is not parallel. For each of the six MULTEXT-EAST languages, the comparable corpus includes two subsets of approximately 100 000 words each: original fiction, comprising a single novel or excerpts from several novels or short stories, and newspapers (see the table below).

Language	Bulgarian	Czech	Estonian	Hungarian	Romanian	Slovene
Fiction	97 251	81 994	104 435	72 002	164 263	95 373
Newspapers	96 538	90 683	112 003	92 233	27 863	101 749

The data in this table is comparable only in terms of the number and size of the texts. The entire multilingual comparable corpus has been prepared in CES format manually or using ad-hoc tools. The corpus is tagged to the paragraph level with some sub-paragraph mark-up (e.g. abbreviations, dates, names).

The data bellow describe the Bulgarian comparable corpus:

Part	Word occurrences	Distinct words	Distinct MSDs in text	Distinct Ctags in text
Fiction	97 251	17 061	313	129
Newspapers	96 538	20 696	295*	126*

* These data have been obtained via Bulgarian_newspapers_lexicon. The Corpus of Bulgarian newspapers is not annotated with MSDs and Ctags by means of MULTEXT tools.

5. Conclusion

The multilingual resources (lexicons, rules, corpora) that have been developed within MULTEXT-East project are probably among of the most valuable resources available for the CEE languages. The MTE corpus is available for research purposes. Currently, the complete documentation of the project, together with HTML corpus "samplers" is available on the worldwide web (http://nl.ijs.si/ME). TELRI also distributes the results of the MULTEXT-East project together with the Plato corpus (developed within the scope of TELRI concerted action) on a two-volume CD-ROM.

Acknowledgements: The project was supported by EU Grant COP 106 (INCO-Copernicus programme). The authors are truly grateful to Nancy Ide and Jean Veronis for the methodological guidance. Special thanks go to Greg Priest-Dorman, Tomaz Erjavec, Heiki-Jan Kaalep, Vladimir Petkevic, Laszlo Tihanyi, and Dan Tufis for the successful cooperation during the project implementation.

References

1. Dimitrova, L., T. Erjavec, N. Ide, H. Kaalep, V. Petkevic, D. Tufis. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. – In: COLING-ACL'98, Montreal, Quebec, Canada, 1998, 315-319.
2. Ide, N., J. Veronis. Multext (multilingual tools and corpora). – In: COLING'94, Kyoto, Japan, 1994, 90-96.
3. Ide, N. Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. – In: First International Conference on Language Resources and Evaluation, LREC'98, Granada, Spain, 1998, 463-470.