Ludmila Dimitrova Radoslav Pavlov

Editors

Mathematical and Computational Linguistics

Jubilee International Conference dedicated to the 30th anniversary of the
Mathematical Linguistics Department
Institute of Mathematics and Informatics - BAS
6 July, 2007, Sofia, Bulgaria

in conj. with the Jubilee International Conference
New Trends in Mathematics and Informatics
dedicated to the 60th anniversary of the IMI - BAS,
6 - 8 July 2007, Sofia, Bulgaria

Proceedings

Programme committee

ISBN 978-954-8986-28-1

Table of Contents

DIGITAL BULGARIAN LANGUAGE RESOURCES - SPECIFICATIONS, CORPORA, LEXICA

Ludmila Dimitrova
Department of Mathematical Linguistics,
Institute of Mathematics and Informatics, BAS
Acad. G. Bonchev str. Bl. 8, 1113 Sofia, Bulgaria
ludmila@cc.bas.bg

Radoslav Pavlov
Department of Mathematical Linguistics,
Institute of Mathematics and Informatics, BAS
Acad. G. Bonchev str. Bl. 8, 1113 Sofia, Bulgaria
radko@cc.bas.bg

Abstract

Our experience with EC projects on language engineering gives us an advantage in the development and extension of digital resources for the Bulgarian language. The Department of Mathematical Linguistics has successfully participated in the EC language technology projects MULTEXT-East (Multilingual Text Tools and Corpora for Central and Eastern European Languages) and CONCEDE (Consortium for Central European Dictionary Encoding) for Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene, as well as for English, as the "hub" language. The first Bulgarian annotated electronic corpus, developed in the MULTEXT-East project, includes two parts: a parallel corpus, based on George Orwell's novel 1984, and a comparable corpus - newspaper excerpts and texts from contemporary Bulgarian literature. The texts of the parallel corpus are produced as a well-structured, lemmatized, CES-corpus. The MULTEXT-East language-specific resources include a lexicon and a set of segmentation and morphological rules and data. Each word-form of the lexicon is associated with the corresponding lemma and its standard lexical description. The digital corpora and lexicon produced mainly serve as input data for experiments with the programs created for processing Western-European languages, but also serve as resources for building lexical databases (LDBs). The CONCEDE project has developed LDBs in a general-purpose document-interchange format for the same seven languages. The project has produced lexical resources that respect the guidelines of the TEI-DWG, and so are compatible with other TEI-conformant resources. Under the CONCEDE project an encoding scheme for lexicographic specifications of the Bulgarian language has been developed according to the standards for electronic dictionary encoding. This encoding scheme serves to create the Bulgarian dictionary in the multilingual LDBs of CONCEDE.

Keywords

digital language resources, electronic corpus, lexical data bases

1. INTRODUCTION

Our experience with EC projects on language engineering gives us an advantage in the development and extension of digital resources for Bulgarian. The Department of Mathematical Linguistics at the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences has successfully participated in the EC language technology projects MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages* and CONCEDE *Consortium for Central European Dictionary Encoding* for Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene, as well as for English, as the "hub" language.

The MULTEXT-East project, as a continuation of the MULTEXT project (Ide and Véronis, (1994)), aims at (1) testing and adaptation of language standards and corpus tools, developed through the MULTEXT, (2) the development of language-specific resources for six new languages and (3) the extension of the annotated multilingual MULTEXT corpus.

MULTEXT *Multilingual Text Tools and Corpora*, one of the largest EU projects in the domain of language engineering, has developed standards and language-specific resources for the encoding and processing of linguistic corpora, as well as tools and corpora embodying these standards. At first, the MULTEXT multilingual corpus included digital corpora in seven Western-European languages: Dutch, English, French, German, Italian, Spanish and Swedish. All results of this project are made freely and publicly available for non-commercial, non-military purposes.

MULTEXT-East (MTE for short) developed digital language resources for six Central and East European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, as well as for English (Dimitrova et al. (1998)). These resources form the MTE multilingual digital database for language engineering research and development. The database contains, for the each of the six CEE languages, morphosyntactic specifications, lexica, and an annotated corpus. The corpus consists of three parts: parallel, comparable and speech-corpus. Thus the MTE database adds six new corpora to the MULTEXT multilingual corpus, (three of which belong to the Slavic group - Bulgarian, Czech and Slovene), which are composed of material comparable to the ones of MULTEXT.

In this way, MTE fulfills its primary goals - to provide examples and test-bed for:

- the applicability of MULTEXT's multilingual tools (especially engine-based tools, alignment software, and multilingual extraction tools) to CEE language corpora;
- the applicability to CEE languages of the Text encoding initiative (TEI) *Guidelines* and MULTEXT's TEI-based corpus markup standard, as well as the MULTEXT-EAGLES (*Expert Advisory Group on Language Engineering Standards*) pan-european lexical specifications and part of speech tagset.

Developed in the frame of the MTE project, Bulgarian language digital resources include morphosyntactic specifications, a lexicon, and a corpus (Dimitrova et al. (2005). The corpus contains two parts which are annotated in accordance with the methodology and requirements of the project - a parallel corpus, based on George Orwell's novel 1984, and a comparable corpus - newspaper excerpts and texts from contemporary Bulgarian literature.

2. BULGARIAN LANGUAGE-SPECIFIC RESOURCES

Morphosyntactic specifications

The MTE morphosyntactic specifications have been developed on the basis of specifications for Western European languages of the EU MULTEXT project (Ide and Veronis, (1994)) and in accordance with the guidelines of the EAGLES. The MTE morphosyntactic specifications contain the list of defined categories – parts of speech (POS): noun, verb, adjective, pronoun, determiner, article, adverb, adposition, conjunction, numeral, interjection, residual, abbreviation, particle. A table of attribute-values is defined for each category, so that each language gets its characteristic features reflectred (Pavlov et al. (1997)). Each part of speech is encoded by a letter: noun - N, verb - V, adjective - A, pronoun - P, determiner - D, article - T, adverb - R, adposition - S, conjunction - C, numeral - M, interjection - I, residual - X, abbreviation - Y, particle - Q. Language particular specifications are marked up additionally by **l.s.**

As an example, for each noun, the following 11 attributes were determined to characterise it as a part of speech, each with its values, harmonized for the six CEE languages and English:
POS: N;
Type: common (c), proper (p);
Gender: masculine (m), feminine (f), neuter (n);
Number: singular (s), plural (p), dual (d), l.s. count (t) *Bulgarian*;
Case: **nominative** (n), genitive (g), dative (d), accusative (a), **vocative** (v), locative (l), instrumental (i), l.s. direct (r) *Romanian*, l.s. oblique (o) *Romanian*, l.s. partitive (1) *Estonian*, illative (x), inessive (2), elative (e), allative (t), adessive (3), ablative (b), l.s. translative (4) *Estonian*, terminative (9), essive (w), l.s. abessive (5) *Estonian*, l.s. komitative (k) *Estonian*, l.s. aditive (7) *Estonian*, l.s. temporalis (m) *Hungarian*, l.s. causalis (c)

Hungarian, l.s. sublative (s) *Hungarian*, l.s. delative (h) *Hungarian*, l.s. sociative (q) *Hungarian*, l.s. factive (y) *Hungarian*, l.s. superessive (p) *Hungarian*, l.s. distributive (u) *Hungarian*;
Definiteness: no (n), yes (y), l.s. short_article (s) *Bulgarian*, l.s. full_article (f) *Bulgarian*;
Clitic: no (n), yes (y);
Animate: no (n), yes (y);
Owner Number: singular (s), plural (p);
Owner Person: first (1), second (2), third (3);
Owned Number: singular (s), plural (p).

The morphosyntactic specifications have been used in the encoding of the word-form lexica of the project.

On the basis of these language specifications, standard lexical descriptions for Bulgarian were determined. The first character of a morphosyntactic description, always letter, encodes the part of speech, e.g., **N for a *noun* or A for an *adjective*** (the full POS-encoding is available above). The characters following the POS-encoding give the values of the position determined attributes. The specifications define, for each part of speech, its appropriate attributes and their values, encoded by one symbol code. It should be noted that in case a certain attribute is not appropriate (1) for a language, (2) for the particular combination of features, or (3) for the word, this is marked by a hyphen in the attribute's position.

The MTE use the MULTEXT format of lexical description - morphosyntactic description (MSD), which consists of linear strings of characters, representing the morphosyntactic information for each word-form. The string is constructed in the following way:

- the positions of a string of characters are numbered 0, 1, 2, etc.
- the agreed character at position 0 encodes the corresponding part of speech: N for ***noun,*** V for ***verb***, etc. ;
- each character at position 1, 2, n, encodes the value of one attribute (for nouns the attributes are: type, person, gender, number, etc.);
- if an attribute does not apply to the word-form, the corresponding position in the string contains a special marker: "-" (hyphen).

For example, the **MSD**

- **Ncmp2** means POS: noun, Type: common, Gender: masculine, Number: plural, Case: inessive;
- **Ncfs-** means POS: noun, Type: common, Gender: feminine, Number: singular, nocase;
- of the Bulgarian word **картата** is **Ncfs-y** that means POS: noun, Type: common, Gender: feminine, Number: singular, nocase, Definiteness: yes.

The proposed formalism for the MSD is not arbitrary (a MSD contains the full description of a lexical item), but has a clear and concrete aim – to be used for specific applications, incl. corpus annotation. A mapping from the morpho-syntactic information, contained in the lexical description, to a set of corpus tags (used by the POS-disambiguator) is also provided, according to the MULTEXT tagging model. The standard lexical descriptions and their respective corpus tags aim to develop a Bulgarian electronic corpus as well as an electronic dictionary of Bulgarian. The list of MSDs for Bulgarian contains 326 elements.

Bulgarian lexicon

The Bulgarian MTE lexicon covers completely the available texts: George Orwell's novel 1984, newspaper excerpts and texts from contemporary Bulgarian literature, which form Bulgarian corpora. A lexicon is a lexical list, containing 17567 lemmata, needed for use in conjunction with the morphological analyser. Table 1 represents the number of lemmata and entries, distributed according to a POS-characteristic.

Table 1. **Number of lemmata and word forms**

POS	Lemmata	Entries
Nouns (total)	9891	47969
Nouns - masculine	4180	25100
Nouns - feminine	4120	16493
Nouns - neuter	1591	6376
Verbs	4140	226666
Adjectives	2155	19397
Pronouns	92	110
Adverbs	790	790
Adpositions	98	98
Conjunctions	76	76
Numerals	67	67
Interjections	172	172
Particles	86	86
Total	**17567**	**295431**

Each element of the lexicon (one entry per line) contains the following information: the inflected-form (word-form), the corresponding lemma and its standard lexical description (MSD) and has the following form

word-form <TAB> **lemma** <TAB> **morphosyntactic description**

An excerpt from the Bulgarian lexicon follows:

Word-Form	**Lemma**	**MSD**
карикатура	=	Ncfs-n
карикатури	карикатура	Ncfp-n
карикатури	карикатуря	Vmia2s
карикатури	карикатуря	Vmia3s
карикатури	карикатуря	Vmip3s
карикатури	карикатуря	Vmm-2s
карикатурист	=	Ncms-n
картина	=	Ncfs-n
картината	картина	Ncfs-y
картини	картина	Ncfp-n
картините	картина	Ncfp-y
картинна	картинен	A--fs-n
картинната	картинен	A--fs-y

Language-specific data

The Bulgarian language-specific resources, except for morphosyntactic specifications and a lexicon, also include a set of segmentation and morphological rules and data, which are necessary for use with the various annotation tools. Segmentation rules describe the form of sentence boundaries, quotations, numbers, punctuation, capitalization, etc. Morphological rules, needed by the morphological tools, provide exhaustive treatment of inflection and minimal derivation. Each lemma in the lexicon is associated with its part(s) of speech and morphological rules. The language-specific data, so called special tokens, required by the segmenter, includes lists of special tokens (frequent abbreviations and names, titles, patterns for proper names, etc.) with their types. Since some subtools in the segmenter require certain language-specific information in order to accomplish their tasks, then each partner has developed a set of resource files for their language. For maximum flexibility and to retain language-independence, all such information is provided directly to the subtools via external resource files, such as:

tbl.punct.Bg
The file tbl.punct.Bg contains the definition of those characters and character configurations that are to be considered as punctuation. They are defined in a regular expression format and each one is assigned an appropriate class, such as "internal punctuation", "non-breaking punctuation", etc.

tbl.abbrev.Bg
The file tbl.abbrev.Bg contains abbreviations ending with a period for the language in question. It is used by the module of the segmenter called mtsegnabbrev. Some abbreviations are identified as belonging to special classes, such as "title", "initial", etc.

tbl.compound.Bg
The file tbl.compound.Bg contains multi-word units which need to be re-combined. Orthographic words which are separated by blanks are split into separate tokens by one of the segmenter's subtools which is invoked early in the chain; this file indicates when such words should be regarded as a single token comprising a compound word.

3. BULGARIAN ANNOTATED ELECTRONIC CORPUS

MTE is building an annotated multilingual corpus, composed of three major parts: **Parallel Corpus, Comparable Corpus**, and a small **Speech Corpus** of spoken texts in each of the six languages, comprising forty short passages of five thematically connected sentences, each spoken by several native speakers, with phonemic and orthographic transcriptions (Dimitrova et al. (1997).

The multilingual parallel corpus, based on George Orwell's novel "1984" in the English original and the six translations in Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene of the novel, was developed. The parallel corpus is produced as a well-structured, lemmatized, CES-corpus (Ide (1998)). The 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 texts were automatically annotated for tokenization, sentence boundaries, and part of speech annotation, using the project tools, and validated for sentence boundaries and alignment. The alignment between the English version and a translation in each of the six CEE languages ensures six pair-wise alignments. The entire text of the 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, please see **APPENDIX 1**.

Bulgarian parallel corpus
The Bulgarian parallel corpus contains the ***Bulgarian translation*** of Orwell's novel "Nineteen Eighty-Four", includes 86020 words (lexical items, excluding punctuation), 101173 tokens (words and punctuations); the ***CesDOC-encoding*** of the Bulgarian text of the novel (SGML mark-up of the text up to the sentence-level) includes 1322 paragraphs, 6682 sentences; the ***CesANA-encoding*** of the Bulgarian text of the novel (containing word-level morpho-syntactic mark-up), and the ***Bulgarian-English aligned texts*** - the aligned versions in ***CesAlign-encoding***, containing links to the aligned sentences (*see* examples bellow). The ***CesANA-encoding*** for Bulgarian, in addition includes: disambiguated lexical information for the 86020 words of the novel and undisambiguated lexical information for 156002 words. What is more – there are 156002 occurrences of MSDs in the text (Bulgarian MSD are 326) and 242022 occurrences of base or lemma of tokens (which is the total of 86020 words and 156002 occurrences of MSD). The number of occurrences of ctags is 257175. Each word-form is associated with the respective grammatical information and the corresponding lemma which form its standard lexical description. The lexical descriptions for Bulgarian are in accordance with the terminology and the methodology used by the MULTEXT project.

An example of the ***CesANA-encoding*** of the Bulgarian text of "Nineteen Eighty-Four"could be found in **APPENDIX 2**.

The next examples show excerpts of the ***Bulgarian-English aligned texts*** – Bulgarian-English Aligned 1984 Sampler:

1-1 Aligned sentences:

<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.

<Obg.1.1.2.10>Портретът бе нарисуван така, че очите да те следват, накъдето и да се обърнеш.
<Oen.1.1.2.10>It was one of those pictures which are so contrived that the eyes follow you about when you move.

<Obg.1.1.2.11>" Големия брат те наблюдава", гласеше надписът отдолу.
<Oen.1.1.2.11>" Big Brother is watching you", the caption beneath it ran.

1-2 Aligned sentences:

<Obg.1.1.14.8>Стомахът му се сви: началото беше съдбоносно.
<Oen.1.1.15.8>A tremor had gone through his bowels. <Oen.1.1.15.9>To mark the paper was the decisive act.

Bulgarian comparable corpus

For each of the six languages, the multilingual comparable corpus was 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, only in terms of the number and size of texts. The entire multilingual comparable corpus was prepared in CES format, manually or using ad-hoc tools. The Bulgarian comparative corpus includes ***Fiction*** (texts from contemporary Bulgarian literature) and ***Newspapers*** (newspaper excerpts) subsets. The Bulgarian ***Fiction*** and ***Newspapers*** subsets were annotated manually. The data in the table below have been determined with the help of the Bulgarian_fiction and Bulgarian_newspapers lexica (the two subsets of the Bulgarian lexicon).

Part	word occurrences	distinct words	distinct MSDs in text	distinct Ctags in text
Fiction	97251	17061	313	129
Newspapers	96538	20696	295	126

The digital corpora and lexicon produced mainly serve as input data for experiments with the programs created for processing Western-European languages, but also serve as resources for building lexical databases (LDBs). The results of the project were made available first on CD-ROM (in 1998), and then via TRACTOR, (*the TELRI Research Archive of Computational Tools and Resources*). The first Bulgarian electronic corpus is included in the *MTE multilingual corpus* of the MTE project, (http://nl.ijs.si/ME), distributed on CD-ROM by *Trans-European Language Resources Infrastructure* (TELRI) Concerted Action Copernicus 1202, (http://www.ids-mannheim.de/telri/) for research purposes.

4. BULGARIAN LEXICAL DATABASES

The CONCEDE project has developed lexical databases (LDBs) in a general-purpose document-interchange format for the same six MTE CEE languages: 3000-headword lexical databases for Bulgarian, Czech, Estonian, Hungarian, Romanian, and a 500-word one from the English-Slovene dictionary. The project has produced lexical resources that respect the guidelines of the Text Encoding Initiative - Dictionary Working Group (TEI-DWG), and so are compatible with other TEI-conformant resources. In this process, the project validated the guidelines (which were developed primarily with reference to Western European languages) for the sixth CEE-languages.

The initial word lists for selection of headwords and word frequency were obtained from the MTE parallel corpus. The selection of headwords was made after word frequency and word class (POS) were taken into account, and the number of words there were in a given word-class and word-frequency band.

In order to achieve a harmonization of the LDBs according to the principal breakdown of lemmata to POS, the CONCEDE consortium decided on the following proportion: open parts of speech (nouns, verbs, adjectives, adverbs) - no more than 90 %, closed parts of speech (numerals, pronouns,conjunctions, prepositions, particles and interjections) – minimum 10% of the whole set of lemmata chosen. The CONCEDE LDBs was developed in two stages. During the first stage, five 500-headwords samples were created by using the different input formalism Document Type Definition (DTD) for producing well-structured SGML-document. During the second stage these 500-headwords LDBs were harmonized and a universal input formalism was created as a language-neutral, dictionary-neutral framework for the presentation of lexical information - CONCEDE DTD.

Under the CONCEDE project was developed an encoding scheme for lexicographic specifications of the Bulgarian language, according to the standards for electronic dictionary encoding (Dimitrova et al. (2002). This encoding scheme served to create the Bulgarian dictionary in the LDBs of CONCEDE. The choice of dictionary entries follows the method accepted by CONCEDE. The entries are equipped with lexicographic specifications for Bulgarian language in TEI-conformant SGML. The electronic dictionary is based on the Bulgarian Explanatory Dictionary (BED) (Andreychin, L. et al. (1994)).

The Bulgarian 500-headwords samples contain 562 lexical entries from BED, selected in accordance with the CONCEDE consortium agreement. Each entry is represented as a tree-structure. These 562 entries contain information for 591 lemmata, because some of the entries contain more than one lemma, for instance, masculine and feminine forms of some nouns. The chosen entries are devided in the following POS: noun – 200 in number or 33.84% of the Bulgarian sample; verb – 130 or 21.99%; adjective – 74 or 12.52%; adverb – 68 or 11.51% -- total open POS 472 or 79.86%; and numeral – 9 in number or 1.52%; pronoun – 31 or 5.24%; conjunction – 24 or 4.06%; preposition – 21 or 3.55%; particle – 26 or 4.40%; interjection – 8 or 1.35% -- total closed POS 119 or 20.13%. The Bulgarian CONCEDE LDBs developed in the second stage of the project contains 2700 entries. The same proportion among the POS is retained here, as well as in the selection of headwords for the Bulgarian sample. The entries in Bulgarian LDBs retain as much as possible the structure of the original paper dictionary:

The entry in the paper Bulgarian Printed Dictionary:
стен|а *ж.* 1. Отвесна, странична част на здание, помещение; зид. *Зидам стена. Външна стена*. 2. Висока каменна или тухлена ограда. *Фернандес лежи в полята пред стените на Мадрид*. Вапц. 3. Вертикална странична част или ограждаща, вътрешна повърхност на нещо кухо. *Казан с дебели стени. Стени на кръвоносен съд*. ◊ И стените имат уши. Китайска стена - нещо, зад което не може да се проникне. Притискам някого до стената - поставям го натясно, в безизходно положение.

The corresponding entry in the Bulgarian LDBs:
<entry>
<hw>стен|а**</hw>**
<gen>ж.**</gen>**
<struc type="Sense" n="1">
<def>Отвесна, странична част на здание, помещение; зид.**</def>**
<eg><q>Зидам стена.**</q></eg>**
<eg><q>Външна стена.**</q></eg></struc>**
<struc type="Sense" n="2">
<def>Висока каменна или тухлена ограда.**</def>**
<eg><q>Фернандес лежи в полята пред стените на Мадрид.
</q><source>Вапц.</source></eg></struc>
<struc type="Sense" n="3">
<def>Вертикална странична част или ограждаща, вътрешна повърхност на нещо кухо.**</def>**
<eg><q>Казан с дебели стени.**</q></eg>**
<eg><q>Стени на кръвоносен съд.**</q></eg></struc>**
<struc type="Phrases">
<struc type="Phrase" n="1"><orth>И стените имат уши.**</orth></struc>**
<struc type="Phrase" n="2"><orth>Китайска стена.**</orth>**
<def>нещо, зад което не може да се проникне.**</def></struc>**
<struc type="Phrase" n="3"><orth>Притискам някого до стената.**</orth>**
<def>поставям го натясно, в безизходно положение.**</def></struc>**
</struc> </entry>

Finally, an examination was carried out – a validation process of the CONCEDE LDBs, which takes two forms, "formal validation" and "content validation". The formal validation was a matter of ensuring that the databases were valid SGML documents and for the Bulgarian LDBs has been done by means of a validating SGML-parser. The content validation of the entries required human intervention and therefore was performed manually.

5. CONCLUSIONS

This article briefly introduces the annotated language resources developed for the first time in Bulgarian in the multilingual research projects of the European Commission MULTEXT-East and CONCEDE. In the course of the work for the development of these multilingual resources, a validation for new six European languages (three of which, incl. Bulgarian, belong to the Slavic group) of the TEI-standards and TEI-guidelines (at first developed for Western-European languages) was carried out. The projects have succeeded in providing foundational resources for work in Language Engineering in the six CEE languages, for morphological, grammatical, semantic or other research, or as the basis for development of new applications in sciences and society.

ACKNOWLEDGEMENT

We would like to thank all colleages with whom we worked throughout the years for the development of the multilingual resources: Kiril Simov and Lydia Sinapova (Bulgarian Academy of Sciences, Sofia, Bulgaria),

Vladimir Petkevič (Charles University, Prague, Czech Republic), Heiki-Jan Kaalep (University of Tartu, Tartu, Estonia), Csaba Oravecz, Laszlo Tihanyi, and Tamas Varadi (Hungarian Academy of Sciences, Budapest, Hungary), Dan Tufiş (RACAI, Romanian Academy, Bucharest, Romania), Tomaž Erjavec (JSI, Ljubljana, Slovenia). We would like to direct a special thanks to the coordinators of the projects Jean Véronis (CNRS-France), Roger Evans and Adam Kilgarriff (ITRI, Brighton, UK), and to Nancy Ide (Vassar College, USA) – for the guidance and the expert help.

REFERENCES

Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimır Petkevič, and Dan Tufiş. (1998). Multext_East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of the COLING-ACL'98*, 315-319, Montréal, Québec, Canada.

Dimitrova, Ludmila, Radoslav Pavlov, and Kiril Simov. (2002). The Bulgarian Dictionary in Multilungual Data Bases. *Cybernetics and Information Technologies.* Vol. 2, num. 2, 12-15.

Dimitrova, Ludmila, Radoslav Pavlov, Kiril Simov, and Lydia Sinapova. (2005). Bulgarian MTE Corpus – Structure and Content. *Cybernetics and Information Technologies.* Vol. 5, num. 1, 67-73.

Ludmila Dimitrova, Lydia Sinapova, Vladimir Petkevič, Jana Klımová, Vera Schmiedtová, Heiki-Jan Kaalep, Viire Villandi, Heili Orav, Leho Paldre, Urve Talvik, Kadri Muischnek, Csaba Oravecz, Laszlo Tihanyi, Ştefan Bruda, Călin Diaconu, Lidia Diaconu, Dan Tufiş, Tomaž Erjavec, Miro Romih, and Olga Vukovič. (1997) Sample corpus collection and preparation. MTE Final Report D2.1F, Institute Jožef Stefan, Ljubljana, Slovenia, December 1997. 70pp.

Tomaž Erjavec, Roger Evans, Nancy Ide, and Adam Kilgarriff. (2000) The Concede model for lexical databases. In *Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, 2000. ELRA.

Tomaž Erjavec, Roger Evans, Nancy Ide, and Adam Kilgarriff. (2003) From Machine Readable Dictionaries to Lexical Databases: the Concede Experience. In *Proceedings of the 7th International Conference on Computational Lexicography, COMPLEX'03*, Budapest, Hungary, 2003.

Nancy Ide. (1998) Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First International Conference on Language Resources and Evaluation, LREC'98*, 463-470, Granada, 1998. ELRA. http://www.cs.vassar.edu/CES/.

Nancy Ide and Jean Véronis. (1994). Multext (multilingual tools and corpora). In *Proceedings of the 15th CoLing*, 90-96, Kyoto, 1994.

Radoslav Pavlov, Ludmila Dimitrova, Lydia Sinapova, and Kiril Simov. (1997) Application to Bulgarian. In *Specifications and notation for lexicon encoding*. MTE Final Report D1.1F, Institute Jožef Stefan, Ljubljana, Slovenia, December 1997, 35-47. http://nl.ijs.si/ME/CD/docs/-d11f/.

The first release of the resourses had been published on the second volume of the "East Meets West" CD-ROM and distributed at cost by TELRI.

APPENDIX 1

Here follows **an excerpt from the CesHeader** of the CesANA-encoding (containing word-level morpho-syntactic mark-up) for Bulgarian language, describing the file, the source of the corpus text, the corpus encoding and its revision history:

```
<encodingdesc>
    <projectdesc>
     MTE:
     Multilingual Text Tools and Corpora for Central and Eastern European Languages.
     EU Copernicus Project COP106
    </projectdesc>
    <editorialdecl>
     <transduction>
      In the cesDoc to cesAna conversion, DIV, QUOTE, Q tags and HEAD, POEM, LIST elements have
      been omitted. cesDoc P elements are encoded as PAR, and S as S.
      cesDoc sub-S level tags are omitted: DATE, NAME, ABBR, etc.
     </transduction>
     <segmentation>
      S segmentation same as in cesDoc source (hand-validated).
      TOK segmentation performed with mtseg and manually corrected.
     </segmentation>
    </editorialdecl>
    <tagsdecl>
     <tagusage gi=chunklist occurs=1>Element corresponds to TEXT of the cesDoc source.</tagusage>
     <tagusage gi=chunk occurs=1>Element corresponds to BODY of the cesDoc source. </tagusage>
     <tagusage gi=par occurs=1322>
      Elements correspond to P elements of the cesDoc source.
      The FROM attribute gives the reference to the ID of the corresponding cesDoc P element.
     </tagusage>
     <tagusage gi=s occurs=6682>
      Elements correspond to S elements of the cesDoc source.
      The FROM attribute gives the reference to the ID of the corresponding cesDoc S element.
     </tagusage>
     <tagusage gi=tok occurs=101173>
      Tokens are of TYPE=WORD or PUNCT, with the CLASS attribute giving the mtseg class of the token.
     </tagusage>
     <tagusage gi=orth   occurs=101173>
      Contains the orthography of the token, as found in the cesDoc source.
     </tagusage>
     <tagusage gi=disamb occurs=86020>Contains disambiguated lexical information.</tagusage>
     <tagusage gi=lex    occurs=156002>
      Contains undisambiguated lexical information.
     </tagusage>
     <tagusage gi=base   occurs=242022> Base or lemma of a token.</tagusage>
     <tagusage gi=msd    occurs=156002> Morphosyntactic description of a token.</tagusage>
     <tagusage gi=ctag   occurs=257175> Corpus tag. </tagusage>
    </tagsdecl>
   </encodingdesc>
   <profiledesc>
    <langusage>
     <![ %ONECOMPONENT [ &ISOlang; ]]>
     <language id=ns-bg iso639=bg>
     Newspeak Bulgarian
     </language>
    </langusage>
   </profiledesc>
```

APPENDIX 2

This appendix demonstrates **an excerpt of the CesANA-encoding** of the Bulgarian text of the "Nineteen Eighty-Four", a morphosyntactic annotation of the second sentence in Orwell's novel:

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.*
(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.)

```
<s from='Obg.1.1.1.2'>
    <tok type=WORD from='Obg.1.1.1.2\1'>
     <orth>С</orth>
     <disamb><base>с</base><ctag>SP</ctag></disamb>
     <lex><base>с</base><msd>Sp</msd><ctag>SP</ctag></lex>
    </tok>
    <tok type=WORD from='Obg.1.1.1.2\3'>
     <orth> глава </orth>
     <disamb><base> глава </base><ctag>NCFS-N</ctag></disamb>
     <lex><base> глава </base><msd>Ncfs-n</msd><ctag>NCFS-N</ctag></lex>
    </tok>
    <tok type=PUNCT from='Obg.1.1.1.2\8'>
     <orth>,</orth>
     <ctag>COMMA</ctag>
    </tok>
    <tok type=WORD from='Obg.1.1.1.2\10'>
     <orth> сгушена </orth>
     <disamb><base>сгуша</base><ctag>VMPS-SF</ctag></disamb>
     <lex><base>сгуша</base><msd>Vmps-sfp-n</msd><ctag>VMPS-SF</ctag></lex>
    </tok>
    <tok type=WORD from='Obg.1.1.1.2\18'>
     <orth>между</orth>
     <disamb><base>между</base><ctag>SP</ctag></disamb>
     <lex><base>между</base><msd>Sp</msd><ctag>SP</ctag></lex>
    </tok>
    <tok type=WORD from='Obg.1.1.1.2\24'>
     <orth>раменете</orth>
     <disamb><base>рамо</base><ctag>NCNP-Y</ctag></disamb>
     <lex><base>рамо</base><msd>Ncnp-y</msd><ctag>NCNP-Y</ctag></lex>
    </tok>
    <tok type=PUNCT from='Obg.1.1.1.2\32'>
     <orth>,</orth>
     <ctag>COMMA</ctag>
    </tok>
    <tok type=WORD class=COMP from='Obg.1.1.1.2\34'>
     <orth>за_да</orth>
     <disamb><base>за_да</base><ctag>CS</ctag></disamb>
     <lex><base>за_да</base><msd>Csc</msd><ctag>CS</ctag></lex>
    </tok>
    <tok type=WORD from='Obg.1.1.1.2\40'>
     <orth>се</orth>
     <disamb><base>се</base><ctag>QV</ctag></disamb>
     <lex><base>се</base><msd>Px---a--yp</msd><ctag>PX</ctag></lex>
     <lex><base>се</base><msd>Qvs</msd><ctag>QV</ctag></lex>
    </tok>
    <tok type=WORD from='Obg.1.1.1.2\43'>
     <orth>скрие</orth>
     <disamb><base>скрия</base><ctag>VMIP3S</ctag></disamb>
     <lex><base>скрия</base><msd>Vmip3s</msd><ctag>VMIP3S</ctag></lex>
```

```
</tok>
<tok type=WORD from='Obg.1.1.1.2\49'>
 <orth>от</orth>
 <disamb><base>от</base><ctag>SP</ctag></disamb>
 <lex><base>от</base><msd>Sp</msd><ctag>SP</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\52'>
 <orth>лютия</orth>
 <disamb><base>лют</base><ctag>AMS</ctag></disamb>
 <lex><base>лют</base><msd>A--ms-s</msd><ctag>AMS</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\59'>
 <orth>вятър</orth>
 <disamb><base>вятър</base><ctag>NCMS-N</ctag></disamb>
 <lex><base>вятър</base><msd>Ncms-n</msd><ctag>NCMS-N</ctag></lex>
</tok>
<tok type=PUNCT from='Obg.1.1.1.2\64'>
 <orth>,</orth>
 <ctag>COMMA</ctag>
</tok>
<tok type=WORD class=COMP from='Obg.1.1.1.2\1'>
 <orth>Уинстън_Смит</orth>
 <disamb><base>Уинстън_Смит</base><ctag>NPMS-N</ctag></disamb>
 <lex><base>Уинстън_Смит</base><msd>Npms-n</msd><ctag>NPMS-N</ctag>
 </lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\1'>
 <orth>се</orth>
 <disamb><base>се</base><ctag>QV</ctag></disamb>
 <lex><base>се</base><msd>Px---a--yp</msd><ctag>PX</ctag></lex>
 <lex><base>се</base><msd>Qvs</msd><ctag>QV</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\4'>
 <orth>шмугна</orth>
 <disamb><base>шмугна</base><ctag>VMIA3S</ctag></disamb>
 <lex><base>шмугна</base><msd>Vmia2s</msd><ctag>VMIA2S</ctag></lex>
 <lex><base>шмугна</base><msd>Vmia3s</msd><ctag>VMIA3S</ctag></lex>
 <lex><base>шмугна</base><msd>Vmip1s</msd><ctag>VMIP1S</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\11'>
 <orth>бързо</orth>
 <disamb><base>бързо</base><ctag>RA</ctag></disamb>
 <lex><base>бръз</base><msd>A--ns-n</msd><ctag>ANS</ctag></lex>
 <lex><base>бърз</base><msd>A--ns-n</msd><ctag>ANS</ctag></lex>
 <lex><base>бързо</base><msd>Ra</msd><ctag>RA</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\17'>
 <orth>през</orth>
 <disamb><base>през</base><ctag>SP</ctag></disamb>
 <lex><base>през</base><msd>Sp</msd><ctag>SP</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\23'>
 <orth>остъклените</orth>
 <disamb><base>остъкля</base><ctag>VMPS-P</ctag></disamb>
 <lex><base>остъкля</base><msd>Vmps-p-p-y</msd><ctag>VMPS-P</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\35'>
 <orth>врати</orth>
 <disamb><base>врата</base><ctag>NCFP-N</ctag></disamb>
 <lex><base>врата</base><msd>Ncfp-n</msd><ctag>NCFP-N</ctag></lex>
</tok>
```

<tok type=WORD from='Obg.1.1.1.2\41'>
 <orth>на</orth>
 <disamb><base>на</base><ctag>SP</ctag></disamb>
 <lex><base>на</base><msd>Qgs</msd><ctag>QG</ctag></lex>
 <lex><base>на</base><msd>Sp</msd><ctag>SP</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\44'>
 <orth>жилищен</orth>
 <disamb><base>жилищен</base><ctag>AMS</ctag></disamb>
 <lex><base>жилищен</base><msd>A--ms-n</msd><ctag>AMS</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\52'>
 <orth> дом </orth>
 <disamb><base>дом</base><ctag>NCMS-N</ctag></disamb>
 <lex><base>дом</base><msd>Ncms-n</msd><ctag>NCMS-N</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\1'>
 <orth>Победа</orth>
 <disamb><base>победа</base><ctag>NCFS-N</ctag></disamb>
 <lex><base>победа</base><msd>Ncfs-n</msd><ctag>NCFS-N</ctag></lex>
</tok>
<tok type=PUNCT from='Obg.1.1.1.2\1'>
 <orth>,</orth>
 <ctag>COMMA</ctag>
</tok>
<tok type=WORD from='Obg.1.1.1.2\4'>
 <orth> но </orth>
 <disamb><base> но </base><ctag>CC</ctag></disamb>
 <lex><base> но </base><msd>Ccs</msd><ctag>CC</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\7'>
 <orth>не</orth>
 <disamb><base>не</base><ctag>QZ</ctag></disamb>
 <lex><base>не</base><msd>Qgs</msd><ctag>QG</ctag></lex>
 <lex><base>не</base><msd>Qzs</msd><ctag>QZ</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\10'>
 <orth>толкова</orth>
 <disamb><base>толкова</base><ctag>RG</ctag>
 </disamb>
 <lex><base>толкова</base><msd>Pd-------q</msd><ctag>PD</ctag></lex>
 <lex><base>толкова</base><msd>Rg</msd><ctag>RG</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\18'>
 <orth> бързо</orth>
 <disamb><base>бързо</base><ctag>RA</ctag></disamb>
 <lex><base>бръз</base><msd>A--ns-n</msd><ctag>ANS</ctag></lex>
 <lex><base>бърз</base><msd>A--ns-n</msd><ctag>ANS</ctag></lex>
 <lex><base>бързо</base><msd>Ra</msd><ctag>RA</ctag></lex>
</tok>
<tok type=PUNCT from='Obg.1.1.1.2\23'>
 <orth>,</orth>
 <ctag>COMMA</ctag>
</tok>
<tok type=WORD from='Obg.1.1.1.2\25'>
 <orth> че </orth>
 <disamb><base> че </base><ctag>QG</ctag></disamb>
 <lex><base> че </base><msd>Ccs</msd><ctag>CC</ctag></lex>
 <lex><base> че </base><msd>Css</msd><ctag>CS</ctag></lex>
 <lex><base> че </base><msd>Qgs</msd><ctag>QG</ctag></lex>
</tok>

<tok type=WORD from='Obg.1.1.1.2\28'>
<orth> да </orth>
<disamb><base> да </base><ctag>QV</ctag></disamb>
<lex><base> да </base><msd>Ccs</msd><ctag>CC</ctag></lex>
<lex><base> да </base><msd>Qgs</msd><ctag>QG</ctag></lex>
<lex><base> да </base><msd>Qvs</msd><ctag>QV</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\31'>
<orth> попречи </orth>
<disamb><base> попреча</base><ctag>VMIP3S</ctag></disamb>
<lex><base> попреча </base><msd>Vmia2s</msd><ctag>VMIA2S</ctag></lex>
<lex><base> попреча </base><msd>Vmia3s</msd><ctag>VMIA3S</ctag></lex>
<lex><base> попреча </base><msd>Vmip3s</msd><ctag>VMIP3S</ctag></lex>
<lex><base> попреча </base><msd>Vmm-2s</msd><ctag>VMM-2S</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\39'>
<orth> на </orth>
<disamb><base> на </base><ctag>SP</ctag></disamb>
<lex><base> на </base><msd>Qgs</msd><ctag>QG</ctag></lex>
<lex><base> на </base><msd>Sp</msd><ctag>SP</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\42'>
<orth> вихрушката </orth>
<disamb><base> вихрушка</base><ctag>NCFS-Y</ctag></disamb>
<lex><base> вихрушка</base><msd>Ncfs-y</msd><ctag>NCFS-Y</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\53'>
<orth> прахоляк </orth>
<disamb><base> прахоляк </base><ctag>NCMS-N</ctag></disamb>
<lex><base> прахоляк </base><msd>Ncms-n</msd><ctag>NCMS-N</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\62'>
<orth>да</orth>
<disamb><base>да</base><ctag>QV</ctag></disamb>
<lex><base>да</base><msd>Ccs</msd><ctag>CC</ctag></lex>
<lex><base>да</base><msd>Qgs</msd><ctag>QG</ctag></lex>
<lex><base>да</base><msd>Qvs</msd><ctag>QV</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\66'>
<orth>нахлуе</orth>
<disamb><base>нахлуя</base><ctag>VMIP3S</ctag></disamb>
<lex><base>нахлуя</base><msd>Vmip3s</msd><ctag>VMIP3S</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\73'>
<orth>с</orth>
<disamb><base>с</base><ctag>SP</ctag></disamb>
<lex><base>с</base><msd>Sp</msd><ctag>SP</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.2\75'>
<orth>него</orth>
<disamb><base>той</base><ctag>PP3</ctag></disamb>
<lex><base>нега</base><msd>Ncf-v</msd><ctag>NCF-V</ctag></lex>
<lex><base>той</base><msd>Pp3msa--n</msd><ctag>PP3</ctag></lex>
<lex><base>то</base><msd>Pp3nsa--n</msd><ctag>PP3</ctag></lex>
</tok>
<tok type=PUNCT from='Obg.1.1.1.2\79'>
<orth>.</orth>
<ctag>PERIOD</ctag>
</tok>
</s>
</par>