MONDILEX: Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources

Russian Academy of Sciences
Institute for Information Transmission Problems
(Kharkevich Institute)

Lexicographic Tools and Techniques

MONDILEX First Open Workshop
Moscow, Russia, 3—4 October, 2008

Proceedings

Leonid Iomdin, Ludmila Dimitrova (Eds.)

The workshop is organized by the project
GA 211938 MONDILEX
Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources
supported by EU FP7 program
Capacities - Research Infrastructures
Design studies for research infrastructures in all S&T fields

Moscow 2008

Le63
УДК 80/81; 004.
ББК 81.1.

Lexicographic Tools and Techniques.
Moscow, IITP RAS, 2008. – 109 p.

The volume contains contributions presented at the First open workshop "Lexicographic tools and techniques", held in Moscow, Russia, on 3—4 October 2008. The workshop is organized by the international project GA 211938 MONDILEX *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources,* Capacities – Research Infrastructures (Design studies for research infrastructures in all S&T fields) EU FP7 programme.

Workshop Programme Committee

Leonid Iomdin (Chairperson), Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

Ludmila Dimitrova (Co-chairperson), Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

Violetta Koseska-Toszewa, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland

Peter Ďurčo, University of St. Cyril and Methodius, Trnava, Slovakia

Radovan Garabík, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

Tomaž Erjavec, Jožef Stefan Institute, Ljubljana, Slovenia

Volodymir Shyrokov, Ukrainian Lingua-Information Fund, National Academy of Sciences of Ukraine, Kiev, Ukraine

Workshop Organising Committee

Leonid Iomdin (Chairperson), Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences

Vyacheslav Dikonov, Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences

Irina Lazurskaya, Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences

Editor of the volume**: Olga Shemanayeva,** Moscow, Russia

ISBN 978-5-9900813-6-9

Contents

On Compatibility of Slavic Language Resources[1]

Ludmila Dimitrova, Radoslav Pavlov
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences, Sofia
ludmila@cc.bas.bg, radko@cc.bas.bg

Abstract
We describe in brief what grid technologies are and how they could contribute to the language technologies, in particular lexicographic activities. Based on our participation in the EC international project MULTEXT-East, we present some aspects of language resource compatibility: unification and standardisation. We underline the importance of the developed harmonised lexical (morphosyntactic) specifications and descriptions of language data in machine-readable form in a common standard encoding format – Corpus Encoding Standard format – for six Central and East European (CEE) languages, as well as the language-independence of the tools employed.

Keywords: language technologies, language resources, grid technologies, electronic corpora, lexicon and dictionaries

Introduction

Applications of language technologies (or natural language processing) have recently been extended in the areas of information research, machine translation, machine learning, speech technology, lexicography, terminological bank servicing, etc.

In a situation of extended applications, language technologies are provoked by new technological decisions (tools) that the information technologies offer recently. A grid, or more precisely, a knowledge grid is such a decision.

What are grid technologies and how could they contribute to the language technologies in particular lexicographic activities?

A grid is a network or collection of distributed computer resources, which are accessible through local or global networks and are presented to the users via an enormous virtual computer system. In a nutshell, a grid is a virtual, dynamically changing organization of structured resources, which are shared among individuals, institutions or systems. Some of the main advantages of the grid technology are: virtual organisation of digital resources; optimized access and enhanced management of these resources; ability to be used worldwide, etc. Knowledge grids offer high-level approaches, techniques and tools for distributed mining and extraction of knowledge from data, processing and accessing of data from the repositories available on the grid, leveraging semantic descriptions of components and data. These functions allow scientists and professionals to compose workflows that integrate data sets and store them, and to create and manage complex knowledge applications. A knowledge grid uses knowledge-based methodologies and technologies to answer much harder questions and to find the appropriate answers in the required form. It joins technologies for data mining, ontologies, intelligent portals, workflow reasoning, etc., for supporting the way knowledge is acquired, used, retrieved, published and maintained.

The relationships between the described features of knowledge grids and lexicographic activities can be briefly formulated as follows:

- Typical knowledge grid objects and language technologies objects (for example, electronic dictionaries and corpora) share some specifications, like: the structural complexity of mono-, bi- and multilingual dictionaries, the great volume of the dictionaries, the internal structure of the dictionaries as a sequence of well-defined tagged-tree lexical entries, etc.
- The knowledge grid provides appropriate services that digital dictionaries require for the coordination and unification of existing digital linguistic resources and for their further cooperative development and enrichment in accordance with recent advances in the field and with international standards, while ensuring their reusability, interoperability (based on open

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

standards and software tools) and openness.

- The knowledge grid allows the creation of an operational structure for the effective communication between the partners and with potential stakeholders, and will support the partners' cooperative efforts to attain the principal objective of the project.
- The possibilities of the knowledge grid technology could provide for the creation of a general lexical database with a rich system of links between forms and meanings of the words; the users could search in any language that already has a digital dictionary.
- The knowledge grid provides infrastructure for the creation and support of a network of high-quality multi-language resources. Many digital lexicographic resources, developed by different research groups or scientists, could be active on the same shared knowledge grid resources at the same time. The research groups could create in collaboration, regardless of distance and time, new digital lexicographic resources that could meet the requirements of the current information space.
- The lexicographic resources (file archives or databases) can be of very different nature, but they must have a standard description, be presented in a standard form in order to be used by standard software tools. This means that the knowledge grid-based infrastructure will support and manage a network of shared resources (e.g., archives, repositories, database, and software tools).
- The high power and secure services of knowledge grid will provide computational techniques for solving some digital lexicographic problems, such as interoperability, ontology integration, content-based automatic selection, automatic content source description, resources preservation, etc.

Annotated electronic Bulgarian language resources

The first annotated electronic resources for Bulgarian language were developed during the EC project MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages* (Dimitrova et al. 1998). Here we give a brief account of our work, as participants in this EC project. We believe that the programming tools used in the MULTEXT-East project (MTE for short) and multi-language resources developed represent a good sample of a **research infrastructure**.

The MULTEXT-East is a continuation of MULTEXT project under the INCO-Copernicus programme. Project MULTEXT produced the 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) (Ide, Veronis 1994).

MTE project developed significant language resources for six Central and Eastern European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, and for English. Three of these languages belong to the Slavic language group: Bulgarian, Czech, and Slovene. The MTE electronic lexical resources include **multilingual MTE corpus**, produced as a well-structured and lemmatized *CES-corpus:* in Corpus Encoding Standard (CES) (Ide 1998, http://www.cs.vassar.edu/CES/), and a dataset of **language specific resources** (for Bulgarian see Dimitrova 1998, Dimitrova et al. 2005).

The results of the two multi-language projects MULTEXT and MULTEXT-East – as resources and experience of using the same program tools – show how important the development of **common harmonised lexical specifications** in *CES-format* for different European languages and the **language-independence of the tools** employed are.

MTE multilingual corpus comprises three corpora: *parallel corpus*, based on George Orwell's novel "1984", *comparable corpus* – newspaper excerpts and texts from CEE literatures, and a small *speech corpus*. The texts of the parallel corpus have been produced as well-structured, lemmatized documents in CES-format.

The language specific resources that MTE project developed are:

- Lexical (morphosyntactic) specifications;
- Language-specific data;
- Lexicons.

The texts and the lexicons produced serve as input data for experiments with the tools created for processing Western-European languages in MULTEXT, but also serve as resources for building lexical databases for the six CEE languages (EC project CONCEDE). The MULTEXT tools were implemented under UNIX. They could be distributed in two main types: corpus annotation tools and corpus exploitation tools – *segmenter, morphological analyser, 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 were designed with an engine-based approach where all language-dependent materials are provided as data (in a form of the tables or rules).

MTE language specific resources

In 1995, the Text Encoding Initiative (TEI), an international project, aimed to produce a guide for preparation and exchange of electronic texts for scientific purposes, using the standards for text representation (http://www.tei-c.org/Guidelines/P5/get_p5.xml).

The TEI-group chose SGML, a metalanguage defined in 1986 with an international standard, ISO 8879, because of its important application to language engineering. SGML makes a text available to many different types of processing or using, because it defines the document's contents entirely and independently of the language. Such text serves as a reusable resource for the purposes of many multilingual systems. The TEI-conformant mark-up techniques (SGML), ensures the efficiency of the electronic exchange of information, large corpora, and lingualware between the scientists of linguistic research. The SGML technique was applied to create language specific resources for the 6th CEE language of the project. The MULTEXT methodology (harmonization of the resources and usage of common tools), used in MTE, has provided producing portable uniform SGML multilingual resources.

The resources for texts processing developed in MTE project are language-specific data. These resources are files required by the MULTEXT project tools for segmentation, tagging, and disambiguation. In this way the language independence of the tools was provided. Each partner has developed a set of resource files for their language and a lexicon according to the common specifications in the MULTEXT format.

MTE lexical specifications

The MTE languages use different character sets and the originals of texts contain symbols not present in ASCII. All MTE electronic texts use 8-bit encoding defined in one of the ISO 8859 standards: Bulgarian uses ISO 8859-5 (Cyrillic), Czech, Hungarian, and Romanian, for example – ISO 8859-2 (Latin 2). The free word order and rich inflection of CEE languages (especially three Slavic: Bulgarian, Czech, and Slovene) presented significantly different linguistic problems than do those of Western Europe. For a description of specific languages phenomena, MULTEXT's specifications were enlarged by an addition of new attributes and values for each Central and Eastern European (CEE) language. So at its first phase the MTE project has developed **harmonised** lexical specifications for six CEE languages and for English (Ide, Veronis, Erjavec (Eds.) 1997).

The specifications are presented as sets of attribute-values, with their corresponding codes used to mark them in the lexicons. The features that are shared by all MTE languages (so-called **core features**) were determined. In such manner the **comparability** of the information encoded in the lexicons across the MTE languages was provided. Except these "general properties" the so-called language-specific features were defined, which describe language-specific morphosyntactic phenomena.

Language-specific data

Sets of segmentation and morphological rules and data for use with the various annotation tools were developed. 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. Other language-specific data, the 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. For maximum flexibility and to retain **language independence**, all such information is provided directly to the subtools via **external resource files,** for example, Bulgarian external files are: tbl.punct.Bg, tbl.abbrev.Bg, tbl.compound.Bg, etc.

MTE lexicons

The MTE lexicons have the **standard** form of the MULTEXT lexica.

Each lexicon entry includes the following information: inflected form; lemma; morphological information for this inflected form encoded in its morphosyntactic description:

wordform <TAB> lemma <TAB> MSD

Examples of Bulgarian lexicon entry:

май	=	**Qgs**	**(Particle, general, simple)**
май	**мая**	**Vmm-2s**	**(Verb, main, imperative, 2nd person, singular)**

In fact, Bulgarian MTE lexicons are three and mostly cover the available texts (Dimitrova et al. 2005):

1. Bulgarian translation of G. Orwell's "1984";
2. Bulgarian corpora
   2.1. *Fiction* (two novels)
   2.2. *Newspaper* (excerpts).

The lexicon of Bulgarian translation of G. Orwell's "1984" contains 17567 lemmas and 295431 word-forms for these lemmas.

The table below shows a number of lemmas and word forms in the Bulgarian lexicon:

Part of Speech	Lemmas	Entries
Nouns	9891	47969
(masculine	4180	25100)
(feminine	4120	16493)
(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

The MTE results

The MULTEXT-East project developed three multilingual corpora:

(1) Parallel Corpus,
(2) Comparable Corpus,
(3) Speech Corpus.

There are **four versions** of **MTE parallel corpus**, corresponding to four different levels of annotation.

For Bulgarian these versions (differently encoded documents) are:

- **Original text** – Bulgarian translation of G. Orwell's novel "1984", includes 86020 words (lexical items, excluding punctuation), 101173 tokens (words and punctuations);
- **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;
- **CesANA-encoding**, containing word-level morpho-syntactic mark-up (undisambiguated lexical information for 156002 words, 156002 occurrences of MSD, and disambiguated lexical information for the 86020 words of the novel);
- **CesAlign-encoding:** Bulgarian-English aligned texts, containing links to the aligned sentences.

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

To arrive at the tokenised and tagged document (for example, G. Orwell's "1984" in Bulgarian) 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. lexical information was disambiguated.

At first, the Bulgarian translation of G. Orwell's "1984" was segmented by means of the segmenter MTSeg – a tokenizer. The segmenter MTSeg 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 **Bulgarian cesDoc Orwell**:

```
<p id="Obg.1.1.2">
  ...
   <s id="Obg.1.1.2.10">Портретът бе нарисуван така, че очите да те следват, накъдето и да се обърнеш. </s>
  ...
  </p>
```

At the S (Sentence) 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. WORD is the values used 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.1.1.2.10">
    <tok type=WORD><orth>Портретът</orth></tok>
    <tok type=WORD><orth>бе</orth></tok>
    <tok type=WORD><orth>нарисуван</orth></tok>
    <tok type=WORD><orth> така </orth></tok>
    <tok type=PUNCT><orth>,</orth><ctag>COMMA</ctag></tok>
    <tok type=WORD><orth>че</orth></tok>
    <tok type=WORD><orth>очите</orth></tok>
    <tok type=WORD><orth>да</orth></tok>
    <tok type=WORD><orth>те</orth></tok>
    <tok type=WORD><orth>следват</orth></tok>
    <tok type=PUNCT><orth>,</orth><ctag>COMMA</ctag></tok>
    <tok type=WORD><orth>накъдето</orth></tok>
    <tok type=WORD><orth>и</orth></tok>
    <tok type=WORD><orth>да</orth></tok>
    <tok type=WORD><orth>се</orth></tok>
    <tok type=WORD><orth>обърнеш</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 "портретът бе" in output of MTLex (in English "picture was" – from the tenth sentence of the "1984": "It was one of those pictures which are so contrived that the eyes follow you about when you move.") are:

```
1.1.2.10\1 TOK Портретът портрет\Ncmsf\NCMS-F
1.1.2.10\11' TOK 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 tenth sentence of the Bulgarian translation of "1984" – *Портретът бе нарисуван така, че очите да те следват, накъдето и да се обърнеш.* (In English: *It was one of those pictures which are so contrived that the eyes follow you about when you move.*).

```
<par from='Obg.1.1.1'>
………………………..
<s from='Obg.1.1.2.10'>
   <tok type=WORD from='Obg.1.1.2.10\1'>
     <orth>Портретът</orth>
     <disamb><base>портрет</base><ctag>NCMS-F</ctag></disamb>
     <lex><base>портрет</base><msd>Ncms-f</msd><ctag>NCMS-F</ctag></lex>
   </tok>
   <tok type=WORD from='Obg.1.1.2.10\11'>
     <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.1.1.2.10\14'>
     <orth>нарисуван</orth>
     <disamb><base>нарисувам</base><ctag>VMPS-SM</ctag></disamb>
     <lex><base>нарисувам</base><msd>Vmps-smp-n</msd><ctag>VMPS-SM</ctag></lex>
   </tok>
   <tok type=WORD from='Obg.1.1.2.10\24'>
     <orth>така</orth>
     <disamb><base>така</base><ctag>QG</ctag></disamb>
     <lex><base>така</base><msd>Qgs</msd><ctag>QG</ctag></lex>
     <lex><base>така</base><msd>Rg</msd><ctag>RG</ctag></lex>
   </tok>
   <tok type=PUNCT from='Obg.1.1.2.10\28'>
    <orth>,</orth>
    <ctag>COMMA</ctag>
   </tok>
   <tok type=WORD from='Obg.1.1.2.10\30'>
     <orth>&chcy;&iecy;</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.1.1.2.10\33'>
   <orth>очите</orth>
   <disamb><base>око</base><ctag>NCNP-Y</ctag></disamb>
   <lex><base>око</base><msd>Ncnp-y</msd><ctag>NCNP-Y</ctag></lex>
 </tok>
 <tok type=WORD from='Obg.1.1.2.10\39'>
   <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.1.1.2.10\42'>
   <orth>те</orth>
   <disamb><base>ти</base><ctag>PP2</ctag></disamb>
   <lex><base>ти</base><msd>Pp2-sa--y</msd><ctag>PP2</ctag></lex>
   <lex><base>те</base><msd>Pp3-pn</msd><ctag>PP3</ctag></lex>
   <lex><base>те</base><msd>Qgs</msd><ctag>QG</ctag></lex>
 </tok>
 <tok type=WORD from='Obg.1.1.2.10\45'>
   <orth>следват</orth>
   <disamb><base>следвам</base><ctag>VMIP3P</ctag></disamb>
   <lex><base>следвам</base><msd>Vmip3p</msd><ctag>VMIP3P</ctag></lex>
 </tok>
 <tok type=PUNCT from='Obg.1.1.2.10\52'>
  <orth>,</orth>
  <ctag>COMMA</ctag>
 </tok>
 <tok type=WORD class=COMP from='Obg.1.1.2.10\55'>
   <orth>накъдето_и_да</orth>
   <disamb><base>накъдето_и_да</base><ctag>RG</ctag></disamb>
   <lex><base> накъдето_и_да </base><msd>Rg</msd><ctag>RG</ctag></lex>
 </tok>
 <tok type=WORD from='Obg.1.1.2.10\69'>
   <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.1.1.2.10\72'>
   <orth>обърнеш</orth>
   <disamb><base>обърна</base><ctag>VMIP2S</ctag></disamb>
   <lex><base>обърна</base><msd>Vmip2s</msd><ctag>VMIP2S</ctag></lex>
 </tok>
 <tok type=PUNCT from='Obg.1.1.2.10\79'>
  <orth>.</orth>
  <ctag>PERIOD</ctag>
 </tok>
</s>
```

Conclusion

As the above examples show, the compatibility of digital resources in Slavic languages (corpora, lexicons, mono-, bi- and multilingual dictionaries) can be achieved through carrying out two major tasks:

- development of standardised and unified lexical descriptions for Slavic languages to annotate texts and word-forms in corpora; lexicon lines; dictionary entries, headwords, etc.,
- use of language-independent programming tools for processing of annotated in such manner language resources.

The grid technologies give us possibilities to:

- transfer and exchange tools and high-volume data (such as digital corpora and dictionaries)
- process in parallel unified data in different Slavic languages by same tools.

The usage of Slavic language resources annotated with standardised and unified lexical descriptions, and the possibilities offered by grid technologies will help linguists in their work to produce new bi- and multilingual Slavic lexical resources and to offer them to the research, education, business communities and to the wide public.

References

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

Dimitrova L. (1998). Lexical Resource Standards and Bulgarian Language. *In International Journal Information Theories & Applications,* Vol. 5, No. 1. pp. 27-34.

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

Ide, N. and Véronis, J. (1994). Multext (multilingual tools and corpora). In *COLING'94*, Kyoto, Japan, pp. 90-96.

Ide, N., Veronis, J., Erjavec, T. (Eds.) (1997). Specifications and Notation for Lexicon Encoding. MULTEXT-East Deliverable D1.1F, Institute Jozef Stefan, Ljubljana, Slovenia. http://nl.ijs.si/ME/CD/docs/mte-d11f/

Ide, N. (1998) Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First International Conference on Language Resources and Evaluation, LREC'98*, Granada, Spain, pp. 463-470.

http://www.tei-c.org/Guidelines/P5/get_p5.xml

http://www.cs.vassar.edu/CES/