The MULTEXT-East Morphosyntactic Specifications for Slavic Languages

Tomaž Erjavec
Dept. of Intelligent Systems
Jožef Stefan Institute, Ljubljana
tomaz.erjavec@ijs.si

Cvetana Krstev
Faculty of Philology
University of Belgrade
cvetana@matf.bg.ac.yu

Vladimír Petkevič
Faculty of Arts
Charles University, Prague
vladimir.petkevic@ff.cuni.cz

Kiril Simov
Linguistic Modelling Laboratory
Bulgarian Academy of Sciences
kivs@bultreebank.org

Marko Tadić
Department of Linguistics
Zagreb University
marko.tadic@ffzg.hr

Duško Vitas
Faculty of Mathematics
University of Belgrade
vitas@matf.bg.ac.yu

Abstract

Word-level morphosyntactic descriptions, such as "Ncmsn" designating a common masculine singular noun in the nominative, have been developed for all Slavic languages, yet there have been few attempts to arrive at a proposal that would be harmonised across the languages. Standardisation adds to the interchange potential of the resources, making it easier to develop multilingual applications or to evaluate language technology tools across several languages. The process of the harmonisation of morphosyntactic categories, esp. for morphologically rich Slavic languages is also interesting from a language-typological perspective. The EU MULTEXT-East project developed corpora, lexica and tools for seven languages, with the focus being on morphosyntactic data, including formal, EAGLES-based specifications for lexical morphosyntactic descriptions. The specifications were later extended, so that they currently cover nine languages, five from the Slavic family: Bulgarian, Croatian, Czech, Serbian and Slovene. The paper presents these morphosyntactic specifications, giving their background and structure, including the encoding of the tables as TEI feature structures. The five Slavic language specifications are discussed in more depth.

1 Introduction

The mid-nineties saw — to a large extent via EU projects — the rapid development of multilingual language resources and standards for human language technologies. However, while the development of resources, tools, and standards was well on its way for EU languages, there had been no comparable efforts for the languages of Central and Eastern Europe.

The MULTEXT-East project (Multilingual Text Tools and Corpora for Eastern and Central European Languages) was a spin-off of the EU MULTEXT project (Ide and Véronis, 1994); it developed standardised language resources for six languages (Dimitrova et al., 1998): Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, as well as for English, the 'hub' language of the project. The main results of the project were an annotated multilingual corpus (Erjavec and Ide, 1998), comprising a speech corpus, a comparable corpus and a parallel corpus; lexical resources (Ide et al., 1998); and tool resources for the seven languages.

One of the objectives of MULTEXT-East has been to make its resources freely available for research purposes. In the scope of the TELRI concerted action the results of MULTEXT-East have been extended with several new languages. This edition is now available via the TELRI Research Archive of Computational Tools and Resources, at *http://www.tractor.de/*.

Following the TELRI release, the MULTEXT-East resources have been used in a number of studies and experiments, e.g., (Tufiş, 1999; Ha-

jič, 2000; Džeroski et al., 2000). In the course of such work, errors and inconsistencies were discovered in the MULTEXT-East specifications and data, most of which were subsequently corrected. But because this work was done at different sites and in different manners, the encodings of the resources had begun to drift apart.

The EU Copernicus project CONCEDE, Consortium for Central European Dictionary Encoding, which ran from '98 to '00 and comprised most of the same partners as MULTEXT-East, offered the possibility to bring the versions back on a common footing. Although CONCEDE was primarily devoted to machine readable dictionaries and lexical databases (Erjavec et al., 2000), one of its workpackages did consider the integration of the dictionary data with the MULTEXT-East corpus. In the scope of this workpackage, the corrected morphosyntactically annotated corpus was normalised and re-encoded. This release of the MULTEXT-East resources (Erjavec, 2001a; Erjavec, 2001b) contains the revised and expanded morphosyntactic specifications, the revised lexica, and the significantly corrected and re-encoded *1984* corpus.

In Table 1, we give all these connected resources by language, type and release. The ones marked by T belong to the TELRI edition, and those with C to the Concede edition. A special case is the Serbian specification, on which we have started working recently.

The columns distinguish the resource in question: "Other Res." are the multilingual tool specifications and the speech and comparable corpora, "*1984* Doc" refers to the structurally annotated parallel Orwell corpus, and "*1984* Align" to the sentence alignments.

By far the most useful part of the MULTEXT-East project deliverables proved to be the morphosyntactic resources, and these were also taken forward to Concede. These resources are also included in the TELRI edition, but have been since substantially modified and added to.

Producing this linked set of deliverables was also by the most labour intensive part of the project. First, while most MULTEXT-East languages had pre-existing morphological lexica and annotations, these had to be 7-way harmonised according to the common specifications, a huge task given not only the diversity of languages but also of linguistic practices. Furthermore, a morphosyntactically annotated corpus of 100,000 words was, for most of the languages, the first such resource to be made. This meant that the annotation had to be done largely manually, and that the corpus annotation process fed back into the lexica and specifications, through a series of revisions.

The morphosyntactic resources consist of three layers, listed in order of abstraction:

1. *1984* MSD: the morphosyntactically annotated *1984* corpus, where each word is assigned its context-disambiguated MSD and lemma, e.g.,
   ```
   <w ana="Pp3ns" lemma="it">It<
   <w ana="Vmis3s" lemma="be">wa
   <w ana="Di" lemma="a">a</w>
   ```
2. MSD Lexicons: the morphosyntactic lexicons, which contain the full inflectional paradigms of a superset of the lemmas that appear in the *1984* corpus. Each entry gives the word-form, its lemma and MSD, e.g.,
   ```
   walk    =       Ncns
   walks   walk    Ncnp
   ```
3. **MSD Specs**: the morphosyntactic specifications, which are the topic of this paper. They set out the grammar of valid morphosyntactic descriptions, MSDs. The specifications determine what, for each language, is a valid MSD and what it means, e.g., `Ncms` → `PoS:Noun, Type:common, Gender:masculine, Number:singular`

To obtain the corpus and lexica, it is necessary to fill out a web-based license agreement, which limits the use of resources to research pruposes. The specifications, however, are freely available on the Web, under *http://nl.ijs.si/ME/*. At the time of writing, the latest version is *V2.1/msd/*

The rest of this paper is structured as follows: Section 2 discusses the structure of the MULTEXT-East morphosyntactic specifications and quantifies them; Section 3 explains the specifications for the Slavic languages; Section 4 turns to the standardisation of the encoding of the specifications in XML/TEI, and Section 5 gives the conclusions and directions for further work.

	Other Res.	*1984* Doc	*1984* Align	*1984* MSD	MSD Lexicon	**MSD Specs**
English	T	T	T/C	C	C	**C**
Romanian	T	T	T/C	C	C	C
Slovene	T	T	T/C	C	C	**C**
Czech	T	T	T/C	C	C	**C**
Bulgarian	T	T	T/C	-	C	**C**
Estonian	T	T	T/C	C	C	C
Hungarian	T	T	T/C	C	C	C
Latvian	-	T	T	-	-	-
Lithuanian	-	T	T	-	-	-
Serbian	-	T	T	-	-	**V2.1**
Russian	-	T	-	-	-	-
Croatian	-	-	-	-	-	**C**

Table 1: The MULTEXT-East Resources: **T**ELRI edition (V1); **C**oncede edition (V2)

2 The Morphosyntactic Specifications

The MULTEXT-East morphosyntactic specifications give the syntax and semantics of the morphosyntactic descriptions (MSDs) used in the lexica and corpora. The specifications have been developed in the formalism and on the basis of specifications for six Western European languages of the EU MULTEXT project (Ide and Véronis, 1994) and in cooperation with EAGLES, the Expert Advisory Group on Language Engineering Standards.

Originally, these specifications were released as a report of the MULTEXT-East project but have, in the CONCEDE release (Erjavec (ed.), 2001), been significantly revised. The format of the report has been unified and structured in a more detailed manner (thus leading to an easily navigable HTML version), the formal specifications for some languages have been modified. The specifications have, in the CONCEDE release also gained a new language, Croatian, and we have recently also added Serbian to the Specifications.

Technically, the specifications are a LaTeX document, with derived Postscript, PDF and HTML renderings, where the common tables are plain ASCII in a strictly defined format. As will be seen in Section 4, we have converted these latter into a TEI/XML encoding.

The MULTEXT-East morphosyntactic specifications have the following structure: (1) introductory matter; (2) the common specification; and (3) a language particular section for each language.

2.1 The Common Part

The common part of the specifications first defines the parts of speech and their codes; MULTEXT-East distinguishes the following, where not all PoS are used for all languages: 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), and Particle (Q).

The common part of the specifications then gives, for each category, a table defining the attributes appropriate for the category, the values defined for these attributes, and one-letter codes to identify the values. They also define which languages distinguish each attribute-value pair. To illustrate, a part of the verb table is given in Table 2.

The morphosyntactic descriptions, MSDs, are structured and more detailed than is commonly the case for part-of-speech tags; they are compact string representations of a simplified kind of feature structures. The first letter of a MSD encodes the part of speech, e.g., Noun or Adjective. The letters following the PoS give the values of the position determined attributes. The specifications define, for each part of speech, its appropriate attributes, their values and one-letter codes. So, for example, the `Ncmpi` MSD expands to `PoS:Noun, Type:common, Gender:masculine, Number:plural,`

```
Verb (V)

15 Positions

**** **** **** **** **** **** **** ---- ---- ---- ---- ---- ---- ---- ----
PoS  Type VFrm Tens Pers Numb Gend Voic Neg  Def  Cltc Case Anim Clt2 Aspt
**** **** **** **** **** **** **** ---- ---- ---- ---- ---- ---- ---- ----

= ============== ============== =  EN  RO  SL  CS  BG  ET  HU  HR  SR
P ATT            VAL            C  x   x   x   x   x   x   x   x   x
= ============== ============== =
1 Type           main           m  x   x   x   x   x   x   x   x   x
                 auxiliary      a  x   x   x   x   x   x   x   x   x
                 modal          o  x   x   x   x       x       x   x
                 copula         c      x   x   x               x   x
                 base           b  x
- -------------- -------------- -
2 VForm          indicative     i  x   x   x   x   x   x   x   x   x
                 subjunctive    s      x
                 imperative     m      x   x   x   x   x   x   x   x
                 conditional    c  x       x   x       x   x   x   x
                 infinitive     n  x   x   x   x       x   x   x   x
                 participle     p  x   x   x   x   x   x       x   x
                 gerund         g      x           x   x
                 supine         u          x           x
                 transgressive  t              x
                 quotative      q                      x
- -------------- -------------- -
...
```

Table 2: The Verb Common Table from the Specifications

`Case:instrumental`. 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 in question, this is marked by a hyphen in the attribute's position. Slovene verbs in the indicative, for example, are not marked for gender or voice, hence the two hyphens in `Vcip3s--n`.

The common part of the specifications further contains two sections, the first giving the complete list of values with their codes and the attributes they belong to, and the second listing all the attributes, which PoS they belong to, and — with the less familiar ones — what their meaning is.

In the context of the common tables, we should mention the Perl script `mtems-expand`, which is — along with various other useful programs — included in the public *msd/bin* directory. The program parses the common tables of the specification, and is then able to either check the validity of any given MSD or expand the MSD into a more readable format.

In Table 3 we quantify the specifications and give, for each language and part-of-speech, the number of attributes and attribute-value pairs defined in the specification. Where a language does not use a PoS, that is marked by a hyphen, while if a PoS is used, but defines no attributes, the value of zero is given.

The numbers in Table 3 give an idea of the "weight" of the MSDs for each particular language and PoS. As can be seen in the bottom right-hand corner, the full number of defined attributes is over one hundred with almost five hundred attribute-value pairs; the Slavic languages define the most categories, and the Pronoun is the most complex PoS.

2.2 Language Particular Sections

In the specifications, the common part is followed by dedicated sections for each particular language, and we turn to these next. The structure of these sections can — although need not — be in itself quite complex. Maximally they contain, for each PoS category, the following parts: (1) attribute-value table with notes; (2) allowed combinations

PoS	**en**	**ro**	**cs**	**sl**	**hr**	**sr**	**bg**	**et**	**hu**	Σ
N	3/7	5/14	5/17	5/16	5/16	5/17	5/14	3/19	7/34	10/54
V	5/15	7/24	10/29	9/28	8/27	8/28	8/24	8/28	6/16	14/52
A	2/4	6/16	7/22	7/23	7/21	7/23	3/9	3/20	8/37	12/61
P	8/23	8/29	12/39	11/40	11/35	10/37	8/30	4/29	7/42	17/88
R	2/7	3/11	2/4	2/5	2/4	2/8	1/2	0	4/13	6/24
S	1/2	4/8	3/8	3/8	3/8	3/8	1/1	1/2	1/1	4/11
C	1/4	5/12	3/7	2/4	2/4	3/8	2/4	1/2	2/6	7/21
M	1/2	6/20	7/29	7/23	6/21	6/20	5/16	4/22	7/39	12/73
I	0	0	0	0	1/2	1/2	1/2	0	1/2	2/4
Y	0	4/15	0	0	4/13	4/14	0	3/21	0	5/35
Q	-	2/7	0	0	1/4	1/4	2/8	-	-	3/15
D	6/16	8/22	-	-	-	-	-	-	-	10/28
T	-	5/13	-	-	-	-	-	-	1/2	5/13
X	0	0	0	0	0	0	0	-	0	0
Σ	29/80	63/191	49/155	46/147	50/155	50/169	36/110	27/143	43/192	107/479

Table 3: Attribute and attribute-value cardinalities of MSDs

of features with examples; (3) full lists of lexical MSDs with examples and cardinality.

The minimal content of a language section is just (1); these are identical to the common ones, but containing only the relevant pairs for the language. These tables can then be further extended, say with notes and examples, and can even be localised to the language in question.

In addition to the tables, the MULTEXT-East languages also have a section giving the feature co-occurrence restrictions on attribute-value pairs. These tables specify the allowed combinations of attribute-values for each PoS, and give a regular expression grammar of MSDs.

The Combinations Sections are useful in the beginning stages of developing lexica, as they isolate malformed MSDs in the resources. However, it is often easier to operate with simple lists of MSDs, as not all possibilities allowed by combinations actually occur in the language.

That is why some languages have, instead of or in addition to the combinations section an explicit list of valid MSDs per category; these lists can then serve as a "gold standard" MSD set for the language; it should be noted that due to rich inflection, the cardinalities of the Slavic language MSDs can be well over a thousand.

3 The Slavic Languages

In this section we further discuss the specifications for the Slavic languages; in particular, we give the historical context in which they were developed and how they related to other MSD tagsets developed for the five languages.

3.1 Bulgarian

At the time when the MULTEXT-East project started there existed two wide coverage morphological lexica for Bulgarian (Morpho-Assistant, Slovnik), both of which encoded the morphosyntactic features of word forms as lists of attribute-value pairs. On the basis of Morpho-Assistant two tagsets were defined: the Bulgarian part of the EAGLES tagset and the LML tagset. On the basis of Slovnik lexicon also two tagsets were defined – first, the Bulgarian part of the MULTEXT-East tagset, which was then extended and localised to Bulgarian (using Cyrillic letters). The two Bulgarian tagsets – LML and Slovnik – are richer than EAGLES and MULTEXT-East tagsets; for a comparison with the LML tagset and discussion see (Slavcheva, 1997).

For the purposes of the BulTreeBank project (Simov et al., 2002), the Slovnik tagset was adapted by having been converted into a Latin format and modified in several ways: there were in-

troduced separate tags for the auxiliary verbs and a hybrid POS tag referring to family names and adjectives derived from names; the pronoun adverbials were made more fine-grained etc. This tagset is being used for the annotation of the BulTreeBank Text Archive. The lexicon is encoded as a regular grammar within the CLaRK system (Simov et al., 2001).

3.2 Croatian

The Croatian specifications were compiled soon after the MULTEXT-East project ended in 1997, using the project's Final report as the template. These specifications are used in the PoS-tagging and lemmatisation of the Croatian National Corpus (Tadić, 2002). It was also selected for the format of MSDs accompanying word-forms in Croatian Morphological Lexicon (Tadić, 2003) which is conformant with MULTEXT-East lexica.

3.3 Czech

The morphological specifications for Czech were developed exclusively for the MULTEXT-East project but the authors had already had some experience with the first draft of morphological specifications for Czech which is now thoroughly described in (Hajič, 2002). These specifications and the resulting tagset developed by Hajič are nowadays used as a standard for morphological and morphosyntactic annotations of the majority of Czech corpora, especially the 100 million word corpus of synchronic Czech developed within the *Czech National Corpus* project. From the present viewpoint, the MULTEXT-East specifications for Czech can be regarded as a subset of this standard. The formalism of both annotation schemes is similar in that both use positional attributes, the important difference being that in MULTEXT-East the attribute position is PoS-dependent, whereas in the standard specifications each attribute is always identified with a fixed position in the tag string.

Among the Czech morphologically annotated corpora, only the Czech translation of *1984* is annotated by the MULTEXT-East specifications. The MULTEXT-East annotation of this corpus was mapped to the standard annotation, i.e., both *1984* corpora differing only in the tagsets used can now be accessed – both are included in the Czech National Corpus.

3.4 Serbian

The Serbian language did not have its representative either in the MULTEXT-East project nor in Concede. The researchers from the Faculty of Mathematics, however, participated in both the TELRI-I and TELRI-II concerted actions. One of the results of this participation was the Serbian *1984* Doc corpus, but the morphosyntactic specification, lexicon and MSD tagged *1984* were not produced.

Independently of these European projects, the same team was working on the production of a Serbian morphological lexicon (Duško Vitas and Cvetana Krstev, 2001) in the format of the INTEX system, which is based on the technology of finite-state transducers (Silberztein, 2000).

The team from the Faculty of Belgrade plans to convert its INTEX lexicon to a MSD-type lexicon. It is to be expected that Serbian MSDs will not differ much from the Croatian ones, as Serbian and Croatian are at the morphological level very similar. The combination of features and lexicon itself will exhibit more differences. A further plan is to produce the annotated version of *1984* that will also be used in the scope of BalkaNet project for the validation of the Serbian WordNet being produced, along with the other languages involved in both MULTEXT-East and BalkaNet, i.e., Czech, Bulgarian and Romanian.

3.5 Slovene

The first version of the Slovene specifications was produced in the scope of the MULTEXT-East project. The second version of the guidelines was produced for the 100 million word FIDA Slovene reference corpus, (Krek et al., 1998). Here the specifications were revised and localised. In particular, all the PoS, attribute, and value names, as well as value codes have been translated into Slovene; the Slovene MSDs are used in the FIDA corpus. The localisation is achieved by extending the tables with additional columns, giving the translation of the symbol(s) and code.

The FIDA MSD specifications were subsequently harmonised with the common MULTEXT-East tables and then released in the context of

CONCEDE; since then they have been used in a number of other corpus projects.

4 The TEI encoding

As has been mentioned, the complete specifications are written in LaTeX, where the common tables are plain ASCII in a strictly defined format. This, over time, has proved to be a good choice, as the format had to be portable and durable, as well as useful for further processing. While we did write several Perl scripts to process or use the common tables, their structure and that of other parts of the specifications (e.g., the combinations) are still quite implicit, and writing a parsing program is not trivial.

For re-use it would certainly be beneficial if the specifications were converted into a standard interchange format, with the obvious choice being XML. As the MULTEXT-East corpus is already encoded in TEI (Sperberg-McQueen and Burnard, 2002), we pursued the option of using already existing TEI tag-sets to encode (parts of) the specifications.

We have defined the MSD IDs in a TEI feature-value library. Additionally, we have also taken the common tables of the specifications and converted these to a TEI feature library, and provided a decomposition from the IDs (MSDs) to the attribute-values and their names.

First, we needed to define the list of all valid MSDs. This, of course, includes the MSDs used in the corpus, but also the MSDs culled from the lexicons; this list then constitutes the authoritative set of valid MSDs for each particular language, and is also included in the language specific sections of the specification.

The MSDs are then encoded as a feature structure library, ⟨*fsLib*⟩, where each MSD is expressed as a feature structure specifying its *type* (the category, i.e., Part of Speech), the language(s) the MSD is appropriate for, and its decomposition into features. The value of ⟨*feats*⟩ is of type IDREFs, i.e., it contains pointers to the definitions of the attribute/value pairs, e.g., `<fs id="Npmpa" type="Noun" select="cs sl" feats="N1.p N2.m N3.p N4.a"/>`

The attribute/value pair definitions are given in the common tables of the morphosyntactic specifications and are encoded as a TEI feature library, ⟨*fLib*⟩. For each feature we give, apart from its identifier, the languages it is appropriate for and the full name of its attribute, while its value is encoded as the content of the feature, as a symbol with the full name of its *value*, e.g., `<f id="N4.a" select="cs hu sl" name="Case"><sym value="accusative"/>`

In the corpus, both libraries are stored in a dedicated corpus element, together with the TEI header. Eventually, the complete morphosyntactic specifications should be converted from LaTeX to TEI and stored in this element.

5 Conclusions

The paper presented the EAGLES & MULTEXT-based multilingual morphosyntactic specifications, which currently include five Slavic languages. Presented were the MULTEXT-East project deliverables and their various editions, esp. those that deal with morphosyntactic resources. The structure and formats of the specifications were discussed, and the Slavic languages introduced in more depth.

As mentioned, of the current Slavic languages, Croatian and Serbian do not yet have the lexical and corpus resource utilising the MSDs defined in the specifications; we hope to remedy this shortcoming sometime in the future, as only with such resources can we validate, quantify and exemplify the specifications. It should be noted that both languages already have lexica that need only to be converted to MULTEXT-East MSDs but producing the MSD tagged *1984* corpus is more complex; while both languages already have the text in digital form, the manual annotation of 100,000 tokens with MSDs is a labour intensive process.

The format of the specifications makes it quite easy to add new languages, although choosing which attributes and values to use, and which word-forms and lemmas to assign them too is far from simple, not only because of the difference in languages, but also due to different linguistic traditions as well as computational models.

In our further work on the specifications, it would be of course beneficial to add new languages, and also to re-evaluate some current

choices in the specifications. On the encoding side, we would like to move to complete specifications to a full TEI/XML encoding and XSLT processing.

Acknowledgements

The complete lists of contributors and acknowledgements are given in the MULTEXT-East Morphosyntactic Specifications, also in the language particular sections. The authors would like to thank all the people mentioned there. The work on these specifications was supported by EU projects MULTEXT-East, CONCEDE and TELRI-II. The work on the individual languages was further supported by various partners' grants and contracts.

References

Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimír Petkevič, and Dan Tufiş. 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. http://nl.ijs.si/ME/.

Duško Vitas and Cvetana Krstev. 2001. Intex and Slavonic Morphology. In *4es Journées INTEX*, Bordeaux. In print.

Sašo Džeroski, Tomaž Erjavec, and Jakub Zavrel. 2000. Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. In *Second International Conference on Language Resources and Evaluation, LREC'00*, pages 1099–1104, Paris. ELRA.

Tomaž Erjavec and Nancy Ide. 1998. The MULTEXT-East corpus. In *LREC'98*, pages 971–974, Granada. ELRA.

Tomaž Erjavec, Roger Evans, Nancy Ide, and Adam Kilgarriff. 2000. The Concede Model for Lexical Databases. In *LREC'00*, pages 355–362, Paris. ELRA.

Tomaž Erjavec (ed.). 2001. Specifications and Notation for MULTEXT-East Lexicon Encoding. MULTEXT-East Report, Concede Edition D1.1F/Concede, Jožef Stefan Institute, Ljubljana. http://nl.ijs.si/ME/V2/msd/.

Tomaž Erjavec. 2001a. Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984. In *6th Natural Language Processing Pacific Rim Symposium, NLPRS'01*, pages 487–492, Tokyo.

Tomaž Erjavec. 2001b. The MULTEXT-East Resources Revisited. *ElsNews*, 10(1):3–2.

Jan Hajič. 2000. Morphological Tagging: Data vs. Dictionaries. In *ANLP/NAACL 2000*, pages 94–101, Seattle.

Jan Hajič. 2002. *Disambiguation of Rich Inflection (Computational Morphology of Czech), Vol. 1*. Karolinum Charles University Press, Prague.

Nancy Ide and Jean Véronis. 1994. Multext (multilingual tools and corpora). In *COLING'94*, pages 90–96, Kyoto.

Nancy Ide, Dan Tufiş, and Tomaž Erjavec. 1998. Development and Assessment of Common Lexical Specifications for Six Central and Eastern European Languages. In *LREC'98*, pages 233–240, Granada. ELRA.

Simon Krek, Marko Stabej, Vojko Gorjanc, Tomaž Erjavec, Miro Romih, and Peter Holozan. 1998. FIDA: a Corpus of the Slovene Language. http://www.fida.net/.

Max Silberztein. 2000. *INTEX*. Masson.

Kiril Simov, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov, and Atanas Kiryakov. 2001. CLaRK – an XML-based System for Corpora Development. In *Corpus Linguistics 2001*, pages 558–560, Lancaster, England.

Kiril Simov, Gergana Popova, and Petya Osenova. 2002. HPSG-based syntactic treebank of Bulgarian (BulTreeBank). In Andrew Wilson, Paul Rayson, and Tony McEnery, editors, *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, pages 135–142. Lincom-Europa, Munich.

Milena Slavcheva. 1997. A Comparative Representation of Two Bulgarian Morphosyntactic Tagsets and the EAGLES Encoding Standard. http://www.lml.acad.bg/projects/BG-EUstand/.

C. M. Sperberg-McQueen and Lou Burnard, editors. 2002. *Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines*. The TEI Consortium. http://www.tei-c.org/.

Marko Tadić. 2002. Building the Croatian National Corpus. In *LREC'02*, pages 441–446, Paris. ELRA.

Marko Tadić. 2003. Building the Croatian Morphological Lexicon. In *[this volume]*. ACL.

Dan Tufiş. 1999. Tiered Tagging and Combined Language Model Classifiers. In Jelinek and Noth, editors, *Text, Speech and Dialogue*, number 1692 in LNAI, pages 28–33, Berlin. Springer-Verlag.