MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora

Tomaž Erjavec

Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

The paper presents the fourth, "Mondilex" edition of the MULTEXT-East language resources, a multilingual dataset for language engineering research and development, focused on the morphosyntactic level of linguistic description. This standardised and linked set of resources covers a large number of mainly Central and Eastern European languages and includes the EAGLES-based morphosyntactic specifications; morphosyntactic lexica; and annotated parallel, comparable, and speech corpora. The fourth release of these resources introduces XML-encoded morphosyntactic specifications and adds six new languages, bringing the total to 16: to Bulgarian, Croatian, Czech, Estonian, English, Hungarian, Romanian, Serbian, Slovene, and the Resian dialect of Slovene it adds Macedonian, Persian, Polish, Russian, Slovak, and Ukrainian. This dataset, unique in terms of languages covered and the wealth of encoding, is extensively documented, and freely available for research purposes at *http://nl.ijs.si/ME/V4/*.

1. Introduction

The MULTEXT-East project, (Multilingual Text Tools and Corpora for Eastern and Central European Languages) ran from '95 to '97 and developed standardised language resources for six CEE languages (Dimitrova et al., 1998), as well as for English, the 'hub' language of the project. The main results of the project were morphosyntactic specifications, defining the tagsets for lexical and corpus annotations in a common format, lexical resources and annotated multilingual corpora. In addition to delivering resources, a focus of MULTEXT-East was also the adoption and promotion of encoding standardization. On the one hand, the morphosyntactic annotations and lexica were developed in the formalism used for six Western European languages in the MULTEXT project (Ide and Véronis, 1994), itself based on the EAGLES specifications (*http://www.ilc.cnr.it/EAGLES/home.html*). On the other, all the corpus resources were encoded in SGML, according to the Corpus Encoding Standard.

After the completion of the EU MULTEXT-East project, a number of other projects have helped to keep the MULTEXT-East resources up-to-date (e.g., migrating the corpus from SGML to XML) and enabled the addition of new languages; the last release of the resources was Version 3 (Erjavec, 2004), which contains resources for 10 languages.

The MULTEXT-East resources have been instrumental in advancing the state-of-the-art in language technologies in a number of areas, e.g., part-of-speech tagging (Tufiş, 1999), inductive learning of lemmatisation rules (Erjavec and Džeroski, 2004), and word sense disambiguation (Ide et al., 2002), to mention just a few. The morphosyntactic specifications have also become a de-facto standard for a number of languages, e.g., Romanian, Croatian and Slovene and people form over 100 institutions have requested use of the resources. The success of the resources is mostly due to the fact that they are freely available for research and that they include uniformly encoded basic resources for processing a large number of languages, for which language resources are still (relatively) scarce. As the linguistic markup has been manually validated and tested in practice, the resources could serve as a "gold standard" which enabled other researchers to develop and test their approaches to topics in language processing.

The resources also provided a model which languages lacking available basic linguistic resources, such as tagsets, lexica and annotated corpora could link up to, taking a well-trodden path. This aspect of the resources was unexpected but highly rewarding and gives impetus for continued work on the overall design of the MULTEXT-East resources.

For Version 4 release of the resources, a substantial part of the work was undertaken in the scope of the project Mondilex. Table 1 gives the grid of the "Mondilex" edition languages and resources, together with the version number when they were made. As can be seen in the table, Version 4 adds, to varying degrees, six new languages: Persian (QasemiZadeh and Rahimi, 2006), Macedonian (Ivanovska et al., 2006), Russian (Sharoff et al., 2008), and, via Mondilex, Polish (Kotsyba et al., 2009), Ukrainian (Derzhanski and Kotsyba, 2009), and Slovak (Garabík et al., 2009). Furthermore, the resources for three languages have been corrected or enhanced.

Version 4 also brings with it also some structural enhancements. The resources have been (re-)coded to be compliant with the current edition of the Text Encoding Initiative Guidelines, TEI P5 (TEI Consortium, 2007). For corpus resources this was a largely automatic procedure, unlike that for the morphosyntactic specifications. Up to and including Version 3 these specifications have been a LaTeX file suitable for printing and conversion to HTML but not for direct processing over MSDs and features. For Version 4 the specifications have been up-translated to (or authored in) XML, making it much simpler to validate and process them automatically. The uniform XML encoding integrates the specifications, lexica and the morphosyntactically annotated corpus, making it possible to easily move between different representations of the same data. This development was largely tied to the development of the JOS monolingual Slovene corpora (Erjavec et al., 2010), where the MSD

Language	Family	MSD Specs	MSD Lexicon	1984 MSD	1984 Alignment	1984 Corpus	Comparable Corpus	Speech corpus
English	Germanic	V1	V1	V1	N/A	V1	V1	-
Romanian	Romance	V1	V1	V1	V1	V1	V1	V1
Polish	West Slavic	V4	V4	V4	-	-	-	-
Czech	West Slavic	V1	V1	V1	V1	V1	V1	-
Slovak	West Slavic	V4	V4	V4	-	-	-	-
Slovene	South West Slavic	V1/V4	V1/V4	V1/V4	V1	V1	V1	V1
Resian	dialect of Slovene	V3/V4	V4	-	-	-	-	-
Croatian	South West Slavic	V2	-	-	-	-	-	-
Serbian	South West Slavic	V2	V2/V4	V3	V3	V2	-	-
Russian	East Slavic	V4	V4	-	-	V2	-	-
Ukrainian	East Slavic	V4	V4	-	-	-	-	-
Macedonian	South East Slavic	V4	V4	V4*	V4	-	-	-
Bulgarian	South East Slavic	V1	V1	V1†	V1	V1	V1	-
Persian	Indo-Iranian	V4	V4	V4	-	-	-	-
Estonian	Finno-Ugric	V1	V1	V1	V1	V1	V1	V1
Hungarian	Finno-Ugric	V1/V4	V1/V4	V1	V1	V1	V1	V1

Table 1: The MULTEXT-East resources by language, and the version when the resources were made / substantially updated. **MSD specs** = morphosyntactic specifications; **MSD lexicon** = morphosyntactic lexicon (entry = wordform+lemma+MSD); **1984 MSD** = MSD and lemma annotated "1984" corpus (* = non-disambiguated; † = automatically tagged with reduced tagset only); **1984 alignment** = sentence alignments with English; **1984 corpus** = "1984", extensively annotated with structural information (poems, quoted speech, etc.) 100,000 words per language; **Comparable corpus** = fiction and newspaper structurally annotated corpus, 2 x 100,000 words per language; **Speech corpus** = parallel speech corpus, 200 sentences per language, spoken + text.

annotation followed a modified form of the MULTEXT-East Version 3 MSD tagset, with the intention being that the JOS morphosyntactic specifications are identical to the Slovene specifications of MULTEXT-East Version 4.

2. The Morphosyntactic Specifications

The MULTEXT-East morphosyntactic specifications are a TEI P5 document that provides the definition of the attributes and values used by the various languages for word-level syntactic annotation, i.e., they provide a formal grammar for the morphosyntactic properties of the languages covered. In addition to the formal parts the specifications also contain commentary, bibliography, etc. The MULTEXT-East specifications, following the original MULTEXT proposal, define 12 categories (mostly corresponding to parts-of-speech), each of which then defines its attributes and their values and the languages that each particular attribute-value pair is appropriate for. The morphosyntactic specifications also define the mapping between the feature-structures and morphosyntactic descriptions (MSDs), which are compact strings used in the morphosyntactic lexica and for corpus annotation. For example, they specify that the MSD `Ncms` is equivalent to the feature-structure consisting of the attribute-value pairs `Category = Noun, Type = common, Gender = masculine, Number = singular`. These definitions are expressed in the so called common tables, which also specify for which languages each particular attribute-value pair is appropriate for. Figure 1 gives an example of an attribute definition; it is expressed as a table (itself part of the category table) with the role attribute giving the function of each row and cell.

The second main part of the specifications are the language particular sections. These, in addition to the introductory matter, also contain sections for each category, with the table of attribute-value definitions appropriate for the language. These tables can be automatically derived form the corresponding common tables, but also modified from them, a novelty in Version 4. In particular, the position of the attribute in the MSD can be different from the common tables, leading to much shorter MSDs for particular languages. The tables can also contain localisation information, i.e., the names of the categories,

```
<row role="attribute">
 <cell role="position">2</cell>
 <cell role="name">Formation</cell>
 <cell>
  <table>
   <row role="value">
    <cell role="name">simple</cell>
    <cell role="code">s</cell>
    <cell role="lang">bg</cell>
    <cell role="lang">mk</cell>
    <cell role="lang">ru</cell>
   </row>
   <row role="value">
    <cell role="name">compound</cell>
    <cell role="code">c</cell>
    <cell role="lang">bg</cell>
    <cell role="lang">mk</cell>
    <cell role="lang">ru</cell>
   </row>
  </table>
 </cell>
</row>
```

Figure 1: XML encoding of the common tables. Example gives the definition of the Formation attribute for the Particle category.

attributes, their values and codes in the particular language, in addition to English. This enables expressing the feature-structures and MSDs either in English, or in the language in question. For example, they map the English MSD `Ncmsn` to the Slovene `Somei` i.e., `samostalnik vrsta = občno_ime spol = moški število = ednina sklon = imenovalnik`.

Each language particular section furthermore contains an index containing all the valid MSDs for the language. Each MSD can be accompanied by explicative information, e.g., examples of usage. This index is the authority for the MSD tagset for the language.

An important part of the specifications are the associated XSLT stylesheets, which allow for various transformations over the specifications. They take the specifications as input, usually together with certain command line arguments, and produce either XML, HTML or text output, depending on the stylesheet.

We provide three classes of transformations, the first ones to help in adding a new language to the specifications themselves, the second to transform the specifications into HTML, and the third to validate and transform a list of MSDs. The outputs of the second and third class of transforms are included in the distribution. The specifications rendered in HTML largely follow the formatting of the original LaTeX specifications, while various conversions of the MSD tagsets for each language are provided in a tabular format for easier use. So, for example, that tables give for each MSD a canonical expansion into features, a sort-code for collating the MSDs in "linguist friendly" collation, or localisation equivalents.

3. Lexica

The MULTEXT-East morphosyntactic lexicons have a simple structure, where each lexical entry is composed of three fields: (1) the *word-form*, which is the inflected form of the word, as it appears in the text, modulo sentence-initial capitalisation; (2) the *lemma*, the base-form of the word; and (3) the *MSD*, i.e., the morphosyntactic description, which is in MSDs tagset defined in the appropriate language particular section.

The size of the lexica between the languages varies considerably, as can be seen in Table 2. Similarly varied are the number of distinct MSDs and the proportions between word-forms, lemmas and MSDs. This is not only due to the different properties of the languages, but also to different approaches adopted in their construction (e.g., including full inflectional paradigms of the lemmas or only attested forms), different linguistic traditions, and, in cases, where the lexicon and MSD tagset were converted from a previous resource, constrains imposed by the automatic mapping procedure.

4. The "1984" corpus

A corpus, annotated with context disambiguated MSDs and lemmas, provides the final piece of the "morphosyntactic triad", as it contextually validates the specifications and lexicon, and provides examples of actual usage of the MSDs and lexical items. The parallel corpus included in MULTEXT-East consists of the novel "1984" by G. Orwell and its translations. The complete novel has about 100,000 tokens, although this of course differs between the languages. The corpus words are annotated with hand validated MSDs and lemmas, which makes it suitable for MSD tagging and lemmatisation experiments.

Lg	Entries	Words	Lemmas	MSDs
en	71,784	48,460	27,467	135
ro	428,093	352,186	39,263	616
pl	337,607	174,444	13,601	1,213
cs	184,470	57,246	23,288	1,425
sk	2,461,491	918,668	76,224	1,534
sl	208,002	122,198	41,171	1,902
sl_r	963	789	338	518
sr	412,979	141,509	9,578	950
ru	243,765	225,061	46,913	582
uk	318,547	205,348	15,162	1,239
mk	1,323,715	1,236,542	81,292	765
bg	54,823	40,546	22,620	338
fa	13,006	11,306	6,595	428
et	135,094	89,591	46,933	642
hu	64,003	51,061	28,063	827

Table 2: Counts on MULTEXT-East lexica

This parallel corpus also comes with separate alignment files, which contain hand-validated pair-wise sentence alignments (not necessarily 1-1) between English and the translations. Version 4 adds pair-wise alignments between all the languages, automatically induced from the alignments with English.

The MULTEXT-East parallel corpus is, of course, small and consists of only one text; nevertheless, it provides an interesting experimentation dataset, as there are still few uniformly annotated many-way parallel corpora, esp. of non-Western European languages.

5. Related work

Standardisation of multilingual linguistic features usually proceeds in the scope of large international projects, and while MULTEXT-East has its genesis in such efforts, in particular EAGLES and MULTEXT it has since then proceeded by slowly adding new languages and upgrading its representation formalism, without making any revolutionary changes. In the meantime new initiatives have started with a similar remit, namely to catalogue and standardise morphosyntactic features. We here mention the two most related to the MULTEXT-East, namely GOLD and isoCat, although many other relevant ones exists, mostly as part of the work of ISO TC 37.

GOLD, the General Ontology for Linguistic Description (Farrar and Langendoen, 2003) is and effort to create a freely available domain-specific ontology for linguistic concepts, availabe at *http://linguistics-ontology.org/*. Given that this effort is well advanced, and that (morphosyntactic) terms are extensively documented, also with references to literature, it would be interesting and not too difficult to link the categories, attributes and their values form the MULTEXT-East specifications to GOLD, providing explication of their semantics.

The isoCat Data Category Registry (Kemps-Snijders et al., 2008) is the Web service at *http://www.isocat.org/* implementing the ISO standard 12620:2009 – Terminology and other content and language resources – Specification of data categories and management of a Data Category Registry for language resources. It provides an on-line registry, where, also terms from the domain of morphosyntax can be found. In the longer term it would be interesting to link up MULTEXT-East to isoCat (esp. as isoCat used the definitions of MULTEXT-East V3 in creating its initial registry) but the system and procedure is, for now, rather complex.

6. Conclusions

The paper introduced Version 4 of the MULTEXT-East resources, which are, as their predecessors, freely available for research at the home page of MULTEXT-East, at *http://nl.ijs.si/ME/V4/*.

The resources now cover most Slavic languages, which is esp. important as a) for a number of them, language resources are otherwise still hard to find and b) these languages have many common characteristics, i.e., they exhibit complex behaviour on the morphosyntactic level, and this is the first dataset that enables a qualitative and quantitative comparison between them.

There are a number of obvious areas for further work: to fill in the blanks in the resources for languages that are already part of MULTEXT-East, esp. the annotated "1984" translations and sentence alignments, where these are still lacking. The number of languages covered could also be extended, most obviously by the Western European languages which were already covered by the MULTEXT project, and possibly by other languages that already use MULTEXT-like morphosyntactic specifications.

Acknowledgments

Thanks to Cvetana Krstev, Vladimír Petkevič, Radovan Garabík and Natalia Kotsyba for comments; for all remaining errors only the author is to blame. The work presented in this paper was in part supported by the EU 7FWP under grant agreement 211938 Mondilex "Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and their Digital Resources".

7. References

Ivan A. Derzhanski and Natalia Kotsyba. 2009. Towards a consistent morphological tagset for Slavic languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. In *Mondilex Third Open Workshop: Metalanguage and encoding scheme design for digital lexicography*, pages 9–26, Bratislava, Slovakia. Ľ. Štúr Institute of Linguistic, Slovak Academy of Sciences.

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. ACL.

Tomaž Erjavec and Sašo Džeroski. 2004. Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, 18(1):17–41.

Tomaž Erjavec, Darja Fišer, Simon Krek, and Nina Ledinek. 2010. The JOS linguistically tagged corpus of Slovene. In *Seventh International Conference on Language Resources and Evaluation, LREC'10*, Paris. ELRA.

Tomaž Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC'04*, Paris. ELRA.

Scott Farrar and D. Terence Langendoen. 2003. A linguistic ontology for the Semantic Web. *GLOT International*, 7(3):97–100. http://linguistics-ontology.org/.

Radovan Garabík, Daniela Majchráková, and Ludmila Dimitrova. 2009. Comparing Bulgarian and Slovak Multext-East morphology tagset. In *Mondilex Second Open Workshop: Organization and Development of Digital Lexical Resources*, pages 38–46, Kyiv, Ukraine. Dovira Publishing House.

Nancy Ide and Jean Véronis. 1994. Multext (multilingual tools and corpora). In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 90–96, Kyoto. ACL.

Nancy Ide, Tomaž Erjavec, and Dan Tufiş. 2002. Sense Discrimination with Parallel Corpora. In *Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia, July. ACL.

Aneta Ivanovska, Katerina Zdravkova, Tomaž Erjavec, and Sašo Džeroski. 2006. Learning Rules for Morphological Analysis and Synthesis of Macedonian Nouns, Adjectives and Verbs. In *Proceedings of 5th Slovenian and 1st international Language Technologies Conference*, Jožef Stefan Institute, Ljubljana.

Marz Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2008. ISOcat: Corralling Data Categories in the Wild. In *Sixth International Conference on Language Resources and Evaluation, LREC'08*. Paris.

Natalia Kotsyba, Adam Radziszewski, and Ivan Derzhanski. 2009. Integrating the Polish Language into the MULTEXT-East Family. In *Mondilex Fifth Open Workshop: Research Infrastructure for Digital Lexicography*, Ljubljana, Slovenia. Jožef Stefan Institute.

Behrang QasemiZadeh and Saeed Rahimi. 2006. Persian in MULTEXT-East Framework. In *FinTAL 2006: 5th International Conference on Natural Language Processing*, pages 541–551, Turku, Finland.

Serge Sharoff, Mikhail Kopotev, Tomaž Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a Russian tagset. In *Sixth International Conference on Language Resources and Evaluation, LREC'08*, Paris. ELRA.

TEI Consortium, editor. 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*.

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