MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora

Tomaž Erjavec

Department of Knowledge Technologies
Jožef Stefan Institute
Jamova 31, Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

The paper presents the third edition of the MULTEXT-East language resources, a multilingual dataset for language engineering research and development. 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, defining the features that describe word-level syntactic annotations; medium scale morphosyntactic lexica; and annotated parallel, comparable, and speech corpora. The most important component is the linguistically annotated corpus consisting of Orwell's novel "1984" in the English original and translations. The resources are the results of several EU projects: MULTEXT-East (produced linked resources for Romanian, Slovene, Czech, Bulgarian, Estonian, Hungarian and English), TELRI (added resources for Lithuanian, Croatian, Serbian, and Russian; first release), and CONCEDE (validation, re-encoding; partial re-release). This paper presents the third release of the resources, which brings together the first two, makes them available in TEI P4 XML, and introduces further extensions, e.g., the specification for Resian, a dialect of Slovene. This dataset, unique in terms of languages and the wealth of encoding, is extensively documented, and freely available for research purposes. The paper presents the component resources, reviews some research undertaken on the basis of the first two editions, and discusses future plans.

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 (Armstrong et al., 1998; Ide and Véronis, 1994; EAGLES, 1996). 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); MULTEXT-East 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 project also adapted existing tools and standards to these languages. The main results of the project were lexical resources and an annotated multilingual corpus. The most important resource turned out to be the parallel corpus — heavily annotated with structural and linguistic information — which consists of Orwell's novel "1984" in the English original and translations.

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 (Trans European Language Resources Infrastructure), the results of MULTEXT-East had been extended with several new languages and first released on a CD-ROM, and later through Web download via TRACTOR, the TELRI Research Archive of Computational Tools and Resources.

Following the TELRI release, the MULTEXT-East resources were used in a number of studies and experiments. 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 '98–'00 EU Copernicus project CONCEDE (Consortium for Central European Dictionary Encoding) offered the possibility to bring the versions back on a common footing. Although CONCEDE was primarily devoted to machine readable dictionaries and lexical databases, one of its workpackages did consider the integration of the dictionary data with the MULTEXT-East corpus (Erjavec et al., 2003a). The CONCEDE release contained the revised and expanded morphosyntactic specifications, the revised lexica, and the significantly corrected and re-encoded linguistically annotated "1984" corpus.

In addition to delivering resources per-se, a focus of the MULTEXT-East, TELRI and CONCEDEprojects was also the adoption and promotion of encoding standardisation. On the one hand, the morpholexical annotations and lexica were developed in the formalism of the (EAGLES-based) specifications for six Western European languages of the MULTEXT project (Ide and Véronis, 1994). On the other, in the TELRI edition, all the corpus resources were encoded in SGML, in CES, the Corpus Encoding Standard (Ide, 1998). For the corpus taken forward into the second edition, the Text Encoding Initiative Guidelines were adopted, in particular TEI P3 (Sperberg-McQueen and Burnard, 1999).

This paper details the third version of the MULTEXT-East resources. The main contribution of this release is that it brings together the first two, i.e., offers both the TELRI and CONCEDE versions in one package, where both are now recoded in XML, according to TEI P4 (Sperberg-McQueen and Burnard, 2002), thus enabling them for processing with XML-based tools. This effort was made possible by the participation in the TEI Working Group on SGML to XML migration (Bia et al., 2002). Furthermore, Version 3 adds some new resources, in particular the annotated Orwell for Serbian, and the morphosyntactic specifi-

cation for Resian, a dialect of Slovene.

Version 3 also contains extensive documentation, e.g., navigational HTML pages, which serve to structure and link the resources, and which include the list of participants and indexes to the resource by type and language. While the TEI headers give the most precise and up-to-date information on the corpus components, the documentation also contains a bibliography with copies of the MULTEXT-East project reports (giving details of the resources, e.g., the corpus markup process), published papers, a mirror of the TEI P4 and CES documentation and certain related MULTEXT and EAGLES reports.

The rest of this paper is structured as follows: we first list the "minor" resources, i.e. those that were produced in MULTEXT-East and TELRI but were not carried into the CONCEDE version. As they are typically rather small, their usefulness is limited, still, they might serve as a starting point in various investigations. We next detail the central, CONCEDE part of the resources, which focuses on the word-level syntactic description of the languages. With each resource we list the languages it is available for. We then review research that has been undertaken on the basis of the MULTEXT-East resources, and conclude with the availability of the resources and plans for future developments.

2. The TELRI resources

The resources listed below are further documented in their headers, and also in the original MULTEXT-East project reports, although the information there is no longer current in all aspects.

2.1. Speech Corpus

Languages: Romanian, Slovene, Estonian, Hungarian, (English, Czech, Bulgarian)

MULTEXT-East produced a small corpus of spoken texts taken from the EUROM-1 speech corpus. It comprises the translations (from English) of forty short passages of five thematically connected sentences. For four languages, the texts have also been read, recorded and included in the distribution. The corpus texts contain links to the spoken passages.

2.2. Comparable Corpus

Languages: Romanian, Slovene, Czech, Bulgarian, Estonian, Hungarian

The multilingual comparable corpus contains a fiction part and a news part, where the data is comparable across the languages in terms of the number and size of texts; each of the 12 parts has approx. 100,000 words. The corpus is encoded in TEI P4 and is structurally marked up with over 40 different elements (e.g. ⟨*head*⟩, ⟨*list*⟩, ⟨*byline*⟩, ⟨*foreign*⟩, ⟨*name*⟩, ⟨*q*⟩).

2.3. Structural "1984" and alignments

Languages: English, Romanian, Slovene, Czech, Bulgarian, Estonian, Hungarian, (Latvian), Lithuanian, Serbian, Russian

The multilingual parallel corpus consists of the novel "1984", about 100,000 words in length. The corpus contains extensive headers and markup for document structure, sentences, and various sub-sentence annotations, these similar to the comparable corpus, but better harmonised over languages.

The translations of "1984" have been automatically sentence aligned with the English original, and the alignments hand-validated. The bilingual alignments are valid to xcesAlign.dtd, i.e., are stored not with the primary data but in separate documents, as references to sentence IDs, e.g., ⟨*link xtargets="Osl.1.2.6.6 ; Ocs.1.1.5.6 Ocs.1.1.5.7"/*⟩.

The cesDoc encoded novel then served as the basis for producing the linguistically annotated version. The link between the two is maintained via sentence identifiers.

3. The morphosyntactic resources

By far the most useful part of the MULTEXT-East project deliverables proved to be the morphosyntactic resources, which were first re-released in the CONCEDE edition and consist of three layers:

1. The morphosyntactic specifications, which set out the grammar and vocabulary of valid morphosyntactic descriptions, MSDs. The specifications determine what, for each language, is a valid MSD and what it means, e.g., that *Ncms* is equivalent to *PoS:Noun, Type:common, Gender:masculine, Number:singular*

2. 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.,
   walks walk Ncnp

3. The morphosyntactically annotated "1984" corpus, where each word is assigned its context disambiguated MSD and lemma, e.g.,
   ⟨*w lemma="it" ana="Pp3ns"*⟩It⟨*/w*⟩.

3.1. Morphosyntactic Specifications

Languages: English, Romanian, Slovene, Czech, Bulgarian, Estonian, Hungarian, Serbian, Russian, Croatian, Resian

The syntax and semantics of the morphosyntactic descriptions (MSDs) are given in the MULTEXT-East morphosyntactic specifications, which have been developed in the formalism and on the basis of specifications for six Western European languages of the MULTEXT project 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 been revised for both subsequent releases, and have become, if not a standard, then at least a reference for comparison (Erjavec et al., 2003b).

The complete specifications are structured as a report, and contain introductory chapters, followed by the list of defined categories (parts-of-speech), and then, for each category, a table of attribute-values, and the languages the features are appropriate for. These so called common tables

```
<f id="R0."  select="en ro sl cs bg et hu
   hr sr sl-rozaj" name="PoS">
  <sym value="Adverb"/></f>
<f id="R1.g" select="ro sl cs bg hu hr sr
   sl-rozaj" name="Type">
  <sym value="general"/></f>
<f id="R1.p" select="ro hu" name="Type">
  <sym value="particle"/></f>
```

Figure 2: Morphosyntactic specifications as TEI features

are followed by language particular sections. Each language section is further subdivided, and can contain feature co-occurrence restrictions, examples, notes, and full lists of valid MSDs, as well as localisation information.

The formal core of the specifications resides in the common tables, as they define the features, their codes for MSD representation, and their appropriateness for each language — an example is given in Figure 1.

Technically, the complete specifications are a LaTeX document (with derived Postscript, PDF and HTML renderings), where the common tables are plain ASCII in a strictly defined format. This format is suitable for a printed version, tolerable for one in HTML, and reasonably manageable for modification and addition of new languages. However, it is not suitable for processing needs, in particular to enable smooth manipulation and linking to an XML encoded corpus using the MSDs.

We have therefore implemented a (Perl) conversion of the common tables into XML, using the TEI.fs module, a tagset devoted to encoding feature-structures. This tagset is currently being used as the basis of an evolving ISO standard (currently a Draft International Standard), as part of work of ISO/TC 37/SC4 Language Resource Management.

The XML version of the common tables has one feature library for each category, e.g., ⟨*fLib type="Noun"*⟩. Each feature in such a library is comprised of the identifier, which enables the linkage to corpus MSDs, the name of the attribute, the languages the feature is appropriate for, and the symbol that is its value. Some examples are given in Figure 2.

3.2. Lexicons

Languages: English, Romanian, Slovene, Czech, Bulgarian, Estonian, Hungarian

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*, which is the base-form of the word; where the entry is itself the base-form, the lemma is given as the equal sign; and (3) the *MSD*, i.e., the morphosyntactic description.

To produce the lexica, the token lists of the MULTEXT-East corpus were first fed through morphological analysers in order to produce the lemma list; this list was further extended from the comparable corpus, to arrive at at least 15,000 lemmas – some languages have further extended this, e.g., Romanian to 41,000 lemmas. In the next step, these lemmas were fed back to morphological generators (except for the agglutinative languages) in order to produce

```
<fs id="R" select="et" feats="R0."/>
<fs id="R-p---q" select="en"
    feats="R0. R2.p R6.q"/>
<fs id="Ra" select="bg"
    feats="R0. R1.a"/>
```

Figure 3: MSDs as TEI feature structures

the complete inflected lists, i.e., the full paradigms of the lemmas, which constituted the final lexica of the project.

The MULTEXT-East lexica serve as medium sized morphological lexica for the languages. In addition to explicating the inflectional behaviour of the most common (and, typically, morphologically the most complex) words of the languages, the lexica also serve to establish the definitive set of valid MSDs for the languages.

To serve as a standard registry of MSDs, we converted the lexical MSDs to TEI feature structure libraries, ⟨*fsLib*⟩, one for each category. Here each MSD is expressed as a feature structure specifying its *id*, the language(s) it is appropriate for, and its decomposition into features. Some examples are given in Figure 3.

Both the ⟨*fsLib*⟩s and the ⟨*fLib*⟩s are stored in dedicated ⟨*TEI.2*⟩ element, complete with its TEI header. This document also constitutes a part of the linguistically annotated MULTEXT-East corpus.

3.3. Linguistically annotated "1984"

Languages: English, Romanian, Slovene, Czech, Bulgarian, Estonian, Hungarian, Serbian

This is the centrepiece of the resources, as it contains word level markup, namely context disambiguated lemmas and MSDs, e.g., ⟨*w lemma="it" ana="Pp3ns"*⟩It⟨/*w*⟩ ⟨*w lemma="be" ana="Vmis3s"*⟩was⟨/*w*⟩ ⟨*w lemma="a" ana="Di"*⟩a⟨/*w*⟩ ⟨*w lemma="bright" ana="Af"*⟩bright⟨/*w*⟩ ⟨*w lemma="cold" ana="Afp"*⟩cold⟨/*w*⟩ ...

The corpus is suitable for PoS tagging experiments; Because it was the first such resources for many of the languages it was also the most difficult to produce as the work had to proceed mostly manually.

4. Research based on the resources

The MULTEXT-East resources have served in a number of experiments, some of which are discussed below — a more comprehensive bibliography is available on the MULTEXT-East Web site.

In the area of part-of-speech tagging, the "1984" corpus the first such resource for a number of MULTEXT-East languages, so it is not surprising that a number of experiments investigated aspects of tagging performance on this corpus. An evaluation exercise (Džeroski et al., 2000) compared four state-of-the-art trainable taggers on "1984"; (Hajič, 2000) tested a feature-based tagger on the corpus; and research on tagset reductions was investigated in developing tagging models for Romanian (Tufiş, 1999) and Hungarian (Varadi and Oravecz, 1999). Another strand of research used the corpus to investigate inductive learning of rules for morphological analysis, in order to lemmatise unknown words in a text (Erjavec and Džeroski, 2004).

Adverb (R)

=	==========	==========	=	EN	RO	SL	CS	BG	ET	HU	HR	SR	SL-ROZAJ
P	ATT	VAL	C	x	x	x	x	x	x	x	x	x	x
=	==========	==========	=										
1	Type	general	g		x	x	x	x		x	x	x	x
		particle	p		x					x			
		causal	o							x			

Figure 1: Start of Common Table for Adverb

For word sense disambiguation, the "1984" corpus served as a testbed for experiments (Ide et al., 2002) that used translation equivalents for automatic sense-tagging.

Another kind of use that the corpus has been put to has been the provision of best practises. An example is the encoding of the 100 million word Slovene reference corpus FIDA (Krek et al., 1998), where both the encoding of the corpus and the morphosyntactic descriptions were taken from the Slovene part of MULTEXT-East. The resources had a similar role for Romanian, Estonian and, partially, Hungarian.

5. Conclusions

The paper presented Version 3 of the MULTEXT-East resources. As the resources cover a number of inflectionally rich languages, are interlinked, harmonised, have a standardised encoding, and have been manually validated and tested in practice, they can serve as a "gold standard" dataset for language technology research and development.

While portions of the resources are distributed without any restrictions, the resources as a whole are available free of charge for research purposes only, as this was the condition imposed by some copyright holders of the sources.

Version 3 of the resources can be downloaded from the MULTEXT-East home page, *http://nl.ijs.si/ME/*. Access is enabled by filling out and submitting a Web based agreement, which is modelled after the one used by Edinburgh's Language Technology Group.

Currently, there are no plans to start working on Version 4; rather, the focus will be on the utility of V3, in our own research, and in enabling others to use the resources, by providing maintenance, continuing to support their accessibility and correcting errors.

6. References

Armstrong, S., M. Kempen, D. McKelvie, D. Petitpierre, R. Rapp, and H. Thompson, 1998. Multilingual corpora for cooperation. In *First Intl. Conf. on Language Resources and Evaluation, LREC'98*. Granada: ELRA.

Bauman, S., A. Bia, L. Burnard, T. Erjavec, C. Ruotolo, and S. Schreibman, 2002. Migrating Language Resources from SGML to XML: the Text Encoding Initiative Recommendations. In *Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04*. Paris: ELRA.

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

Džeroski, S., T. Erjavec, and J. Zavrel, 2000. Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. In *Second Intl. Conf. on Language Resources and Evaluation, LREC'00*. Paris: ELRA.

EAGLES, 1996. Expert advisory group on language engineering standards. http://www.ilc.pi.cnr.it/EAGLES/home.html

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

Erjavec, T., R. Evans, N. Ide, and A. Kilgarriff, 2003a. From Machine Readable Dictionaries to Lexical Databases: the Concede Experience. In *Proceedings of the 7th Intl. Conf. on Computational Lexicography, COMPLEX'03*. Budapest, Hungary.

Erjavec, T., C. Krstev, V. Petkevič, K. Simov, M. Tadić, and D. Vitas, 2003b. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*. Budapest.

Hajič, J., 2000. Morphological Tagging: Data vs. Dictionaries. In *ANLP/NAACL 2000*. Seattle.

Ide, N., 1998. Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First Intl. Conf. on Language Resources and Evaluation, LREC'98*. Granada: ELRA. http://www.cs.vassar.edu/CES/

Ide, N., T. Erjavec, and D. Tufiş, 2002. Sense Discrimination with Parallel Corpora. In *Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*. Philadelphia: ACL.

Ide, N. and J. Véronis, 1994. Multext (multilingual tools and corpora). In *Proceedings of the 15th Intl. Conf. on Computational Linguistics*. Kyoto.

Krek, S., M. Stabej, V. Gorjanc, T. Erjavec, M. Romih, and P. Holozan, 1998. FIDA: a Corpus of the Slovene Language. *http://www.fida.net/*

Sperberg-McQueen, C. M. and L. Burnard (eds.), 1999. *Guidelines for Electronic Text Encoding and Interchange, Revised Reprint*. The TEI Consortium.

Sperberg-McQueen, C. M. and L. Burnard (eds.), 2002. *Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines*. The TEI Consortium.

Tufiş, D., 1999. Tiered Tagging and Combined Language Model Classifiers. In Jelinek, F. and E. Noth (eds.), *Text, Speech and Dialogue*, number 1692 in Lecture Notes in Artificial Intelligence. Berlin: Springer-Verlag.

Varadi, T. and C. Oravecz, 1999. Morpho-syntactic Ambiguity and Tagset Design for Hungarian. In *Proceedings of the EACL-99 Workshop on Linguistically Interpreted Corpora (LINC-99)*. Bergen: ACL.