MULTEXT-East Morphosyntactic Specifications: Towards Version 4*

Tomaž Erjavec

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

Abstract. The MULTEXT-East standardised and linked set of language resources covers a large number of mainly Central and Eastern European languages and includes harmonised morphosyntactic resources consisting of the specifications, lexica and a parallel corpus. The MULTEXT-East resources, currently at Version 3, are freely available for research use and have been used in numerous studies connected to language technologies. In this paper we concentrate on MULTEXT-East morphosyntactic specifications, which define the features that describe word-level syntactic annotations, and explain their structure in Version 4, currently work in progress. The V4 specifications are planned to cover at least 13 languages and will be encoded in XML, according to the latest version of the Text Encoding Initiative Guidelines, TEI P5. The new encoding enables more flexible language-particular encodings, localisations of feature names and codes, easy generation of derived formats (HTML, tabular, XML libraries), and simplifies the addition of new languages.

1 Introduction

The MULTEXT-East project, (Multilingual Text Tools and Corpora for Eastern and Central European Languages) was a spin-off of the EU MULTEXT project [14]; MULTEXT-East ran from '95 to '97 and developed standardised language resources for six CEE languages [3], as well as for English, the 'hub' language of the project. The main results of the project were lexical resources and an annotated multilingual corpus, where 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.

In addition to delivering resources, a focus of MULTEXT-East was also the adoption and promotion of encoding standardisation. On the one hand, the morphosyntactic annotations and lexica were developed in the formalism used for six Western European languages in the MULTEXT project, itself based on the EAGLES specifications [5]. On the other, all the corpus resources were encoded in SGML, according to the Corpus Encoding Standard [12] and, later, in XML and TEI, the Text Encoding Initiative Guidelines [19].

One of the objectives of MULTEXT-East has been to make its resources available to the wider research community. The resources are distributed on the Web at *http://nl.ijs.si/ME/*. A portion of the resources is freely available for download or browsing; for the rest, the user has to first fill out a Web-based agreement form restricting the use of resources for research. Apart from the data itself, the distribution also contains extensive documentation.

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 us to add new languages. At the time of writing, the latest publicly released resources are at Version 3 [7].

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 [21], inductive learning of

* The study and preparation of these results have received funding from the EU 7FWP under grant agreement 211938 MONDILEX.

lemmatisation rules [9], and word sense disambiguation [13], to mention just a few. The licensing form has been submitted by over 100 organisations, mostly academia, but also industry.

The success of the resources is mostly due to the fact that they are freely available for research and that they include basic building blocks for processing a significant range of "novel" languages. As the linguistic markup has also been manually validated and tested in practice, the resources can serve as a "gold standard" which enables other researchers to develop and test their approaches to topics in the language processing. The resources also provide a model which languages lacking basic linguistic resources, such as tagsets, lexica and annotated corpora can link-up to, taking a well-trodden path. This aspect of the resources was unexpected but highly rewarding; this steady addition of new languages also gives impetus to continue working on their general improvement.

Since the release of Version 3 the resources have again been expanded and re-encoded, in preparation for Version 4. New languages have been added and the morphosyntactic specifications have been converted from the LaTeX format to XML [8]. A portion of the resources has also been additionally annotated, e.g., for WordNet word-sense disambiguated nouns [13] in the English "1984" and dependency syntactic structures for the Slovenian "1984" [4].

This paper is devoted to one part of the resources, namely the MULTEXT-East morphosyntactic specifications. The specifications are a document that provides the definition of the attributes and values used by the various languages for word-class syntactic annotation, i.e., they provide a formal grammar for the morphosyntactic properties of the languages covered. The MULTEXT-East specifications define 12 categories (parts-of-speech), and approx. 100 different attributes with 500 values.

The morphosyntactic specifications also define the mapping between feature-structures and morphosyntactic descriptions (MSDs), which are compact strings used in the morphosyntactic lexica and for corpus annotation. For example, they specify that MSD `Ncms` is equivalent to the feature-structure consisting of the attribute-value pairs `Category:Noun`, `Type:common`, `Gender:masculine`, `Number:singular`. The specifications furthermore determine which feature-value combinations and MSDs are valid for particular languages. In addition to the formal parts the specifications also contain commentary, bibliography, etc.

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 [11]. They currently cover thirteen languages; Table 1 gives an overview, and for each language also specifies its language family, and which version of the MULTEXT-East resources it first appeared or will appear in. Special mention deserve the languages which still have to make their debut in Version 4, namely Macedonian, Persian, and Russian, and, to an extent, Slovene. The development of the Macedonian specification, lexicon and corpus started in 2004, and the resources have already been used as the data for several experiments in tagger [22] and lemmatiser induction [15]. The Macedonian resources comprise the specifications, lexicon, and corpus, which is, however, not yet morphosyntactically annotated. The development of Persian resources also started in 2004, and they currently comprise the specifications and annotated corpus [17]. The Russian specifications [18] are the latest addition, although the (unannotated) corpus has been available since Version 1. The Russian resources thus still lack a lexicon and annotated corpus, although an automatically annotated corpus and tagging models are available independently at *http://corpus.leeds.ac.uk/mocky/*.

Slovene has been a part of the MULTEXT-East resources from the start, however, in Version 4 we plan to significantly revise the specifications and harmonise the lexicon and corpus with them. The Slovene specifications have been extensively used for corpus annotation, esp. of the Slovene reference corpora Fida and its successor FidaPLUS (*http://www.fidaplus.net/*) and in the course of the years various shortcomings of the original proposal have come to light. A recent Slovene project, JOS (Jezikoslovno označevanje slovenščine / Linguistic Annotation of Slovene, *http://nl.ijs.si/jos/*), devoted to corpus annotation has provided the means to revise the specifications, and use them as the basis to (semi)manually annotate two corpora of Slovene [10]. The development of these "JOS" specifications, has, to a large extent, also served as the testing ground for the new MULTEXT-East specifications. In Version 4 we plan to incorporate the JOS specifications into MULTEXT-East.

Language name	Language family	Added in
English	Germanic	Version 1
Romanian	Romance	Version 1
Russian	East Slavic	Version 4
Czech	West Slavic	Version 1
Slovene	South West Slavic	Version 1/4
Resian	dialect of Slovene	Version 3
Croatian	South West Slavic	Version 3
Serbian	South West Slavic	Version 2
Macedonian	South East Slavic	Version 4
Bulgarian	South East Slavic	Version 1
Persian	Indo-Iranian	Version 4
Estonian	Finno-Ugric	Version 1
Hungarian	Finno-Ugric	Version 1

Table 1. Languages covered by the morphosyntactic specifications.

The rest of this paper is structured as follows: Section 2 details the XML format of the specifications, Section 3 discusses the associated XSLT stylesheets, Section 4 briefly introduces the MULTEXT-East lexica and annotated corpus, and Section 5 gives some conclusions and directions for further work.

2 The format of the specifications in V4

In this section we give some background in the area of standardisation of multilingual morphosyntactic specifications, and detail their structure and encoding for MULTEXT-East Version 4.

The concepts expressed in MULTEXT-East specifications go back to the EAGLES guidelines from the early '90. The EU project EAGLES, the Expert Advisory Group on Language Engineering Standards, was instrumental for advancing the field of standardisation of language resources in a multilingual setting, and tackled corpora, spoken resources, lexica etc. as well as morphosyntactic descriptions and their specifications [2, 6].

But while the EAGLES compared a large number of proposals and gave general recommendations for encoding morphosyntactic descriptions, it did not provide explicit common specifications for a set of languages which could be mapped into morphosyntactic descriptions as used in lexica and corpora. This did, however, happen in the EU MULTEXT project, where the format of the specifications was concretised [1] for six EU languages (Italian, German, Spanish, French, Dutch, and English). The complete morphosyntactic specifications of MULTEXT were written as a LaTeX document, where the common tables are plain ASCII in a strictly defined format. The MULTEXT proposal also divided the features it defined into "general" and language specific ones. The first are taken to be used by most MULTEXT languages, while the second were those that were felt to be needed to describe the specifics of particular languages and their pre-existing resources.

MULTEXT-East adopted the MULTEXT format, except that it re-defined the language particular features to accommodate the radically different, mainly inflectional properties of the MULTEXT-East languages, and substituted the MULTEXT languages with the MULTEXT-East ones. The two proposals thus cannot be trivially combined, as they share only a subset of the attributes.

The complete MULTEXT-East morphosyntactic specifications consist of the following parts:

1. introductory matter: preface, background, organisation of the proposal, bibliography
2. common part: attribute-value tables for each category with notes
3. language particular parts for each language

The MULTEXT specifications, in particular, the attribute-value tables of the common part, should be interpreted as defining feature-structures, a well-known linguistic representation formalism, where a feature-structure consists of a set of attribute-value pairs. The common tables thus correspond to the definition of attribute- value pairs (e.g., that there exists, for Nouns, an attribute `Type`, which can have the values `common` or `proper`), while an MSD corresponds to a fully-specified feature-structure. But in MULTEXT there was no automatic way (piece of software) provided for converting the MSDs to feature-structures or vice-versa, or for checking the consistency of the specifications. For this reason MULTEXT-East soon developed a (Perl) program„ which could expand, on the basis of the common tables in the specifications, MSDs into a plain text feature-structures or check the validity of an MSD for a given language.

Having the document formatted in LaTeX and the formal parts written as ASCII tables had the virtue of simplicity but was problematic for at least two reasons. As mentioned, ad hoc programs were needed to validate MSDs against the specifications, or to internally validate the specifications. As the years passed, it was also becoming increasingly difficult to add new languages in a controlled fashion, due to the brittleness of the plain text format, and to the inter-dependencies and redundancy between the tables. What was needed was a formal specification for the tables that would enable their validation, extension, rendering on the Web or paper, or conversions into other formats.

2.1 Using the TEI

The Text Encoding Initiative *http://http://www.tei-c.org/* is an international consortium, whose primary function is to maintain the TEI Guidelines, which set out a vocabulary of elements useful for describing text for scholarly purposes. The Guidelines use XML encoding and are written as a set of XML schemas (element grammars) with accompanying documentation. In MULTEXT-East V3 we used Version P4 for encoding of the corpora, while in V4 we use of the most recent published version, TEI P5 [20].

The are a number of advantages of using TEI for encoding. TEI documents are written in XML, which brings with it the possibility of validation of document structure, a wealth of supporting software and related standards. Of these, the most important is the XML transformation language, XSLT, which allows writing scripts (stylesheets) that transform XML documents into other, differently structured XML documents, or into HTML as well as, indirectly, into a printable version in, say, PDF. The XSLT standard is nowadays generally supported, e.g., we find it implemented in most Web browsers. The MULTEXT-East specifications come with a number of XSLT transforms, which help in authoring or displaying the specifications; they are further discussed in Section 3.

TEI is also general enough to encode the non-normative parts of the specifications, e.g., the introductions, notes, references, etc. The TEI also provides, amongst other software, a sophisticated set of XSLT stylesheets and associated components for converting TEI documents into HTML and PDF. These stylesheets, developed by Sebastian Rahtz and freely available via the TEI homepage, cover a large number of TEI elements, and also perform tasks such as generating the table of contents, splitting (large) TEI documents into several HTML files (while preserving cross-links), giving each HTML a project defined header and footer, etc.

Finally, the MULTEXT-East parallel and MSD annotated corpus was already encoded in TEI; by encoding the specifications in TEI as well, this gives an easy way to directly integrate the corpus with the specifications, leading to simple validation of the corpus annotations or conversion between corpus MSDs and their feature-structure representations. This can be extremely useful for querying the corpus, as it enables e.g., the selection of word tokens based on particular features.

For these reasons the V4 specifications are written in TEI P5, as one XML document (which does not mean they have to be in one file), with the idea that this is the single document which needs to be maintained and to which new languages are added in a controlled fashion. The structure should therefore be amenable to hand editing, minimally redundant, contain as much as possible of structured commentary and references, with the formal parts having a transparent structure.

2.2 The common part of the specifications

This section gives more detail about the structure of the common part of the specifications in TEI. The common part of the specifications contains:

1. A table giving all the languages of the specification. For each language the table also gives its language family, ISO 836 code, and a link to its description in the Ethnologue database.
2. A table giving the (part-of-speech) categories of MULTEXT-East (12) together with their one-letter codes. The derived HTML of the specifications (so called display version) additionally contains the number of attributes defined for each category and which languages distinguish them.
3. For each category, the common table, defining attributes and their values for the category. For attribute they also specify its position in the MSD string, and for each attribute-value pair, a one letter code for the MSD string. For each such pair, the table also lists the languages that the attribute-value is valid for.
4. A table of all defined attributes, with the categories they are defined for, and their position in the MSD string (in display version only, and automatically generated from the XML source).
5. A table of all defined values, with the attribute/categories they are defined for, their code in the MSD string, and the languages that distinguish this attribute-value pair (in display version only, and automatically generated from the XML source).

Figure 1 gives an example from the TEI source, while Figure 2 gives the display view; the latter is, on purpose, quite similar to the tables in MULTEXT-East V3. The master TEI is, however, more logically oriented: the first row defines the category and gives the languages it is appropriate for while the following rows each define an attribute, with the values given in a subordinate table.

2.3 The language particular specifications

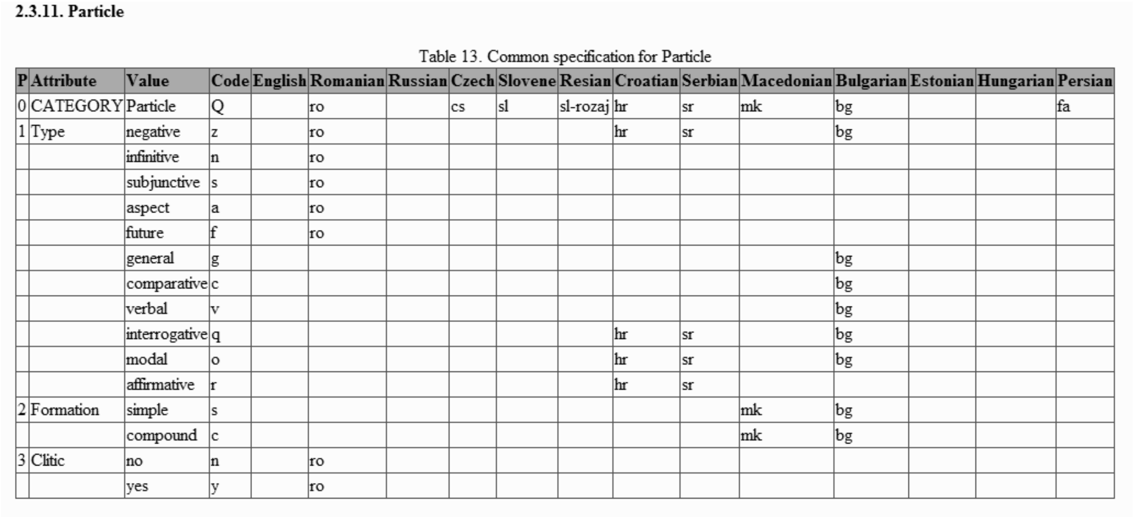
The specifications contain, for each language, also a language particular part. These parts can have a minimal structure, just giving the authors and repeating the common tables, but reduced to the categories and attribute-value pairs that are in fact used by the language. They can also be quite complex and can contain some or all of the following divisions:

- Introductory matter, e.g., language description; background of the language specifications; bibliography.
- Then, for each category:
  - The language particular table, which can be automatically derived form the common table, but also modified from it, as will be further described below. Furthermore, the tables can also contain localisation information, i.e., the names of the categories, attributes, their values and codes in the particular language, in addition to English. This enables keeping the feature-structures and MSDs either in English, or in the language in question.
  - Notes on the category itself or on the attributes and values used.
  - Combinations of attribute-values (feature co-occurrence restrictions), which in a regular-expression-like syntax limit the possible combinations of attribute-values. These restrictions can also contain examples of usage. It should be noted that these combinations have not yet been operationalised, i.e., it is not possible to directly use them to validate MSDs.
  - A list of lexical MSDs, which should contain all the valid MSDs for the category. This is present only in the display view and automatically extracted from the full MSD index.
- The MSD index, which should contain all the valid MSDs for the language. Each MSD can be furthermore accompanied by explicatory information, i.e., its decomposition into feature-values, examples of usage, and its translation. This index is the authority for the MSD set for the language, and is valuable for MSD validation.

As an example of how a language particular table can look in Version 4, we give the JOS table for Slovene Nouns in Figure 3. The table gives identical information as the (Slovene selected) common tables, except that all information is also translated/localised to Slovene.

```
<div type="section" id="msd:Q">
  <head>Particle (Q)</head>
  <table n="mtems-cat">
    <head>Common specifications for Particle</head>
    <row role="type">
      <cell role="position">0</cell>
      <cell role="name">CATEGORY</cell>
      <cell role="value">Particle</cell>
      <cell role="code">Q</cell>
      <cell role="lang">ro</cell>
      <cell role="lang">sl</cell>
      ...
    </row>
    <row role="attribute">
      <cell role="position">1</cell>
      <cell role="name">Type</cell>
      <cell role="status">common</cell>
      <cell>
        <table>
          <row role="value">
            <cell role="name">negative</cell>
            <cell role="code">z</cell>
            <cell role="lang">ro</cell>
            <cell role="lang">bg</cell>
            <cell role="lang">hr</cell>
            <cell role="lang">sr</cell>
          </row>
          <row role="value">
            <cell role="name">infinitive</cell>
            <cell role="code">n</cell>
            <cell role="lang">ro</cell>
          </row>
          <row role="value">
            <cell role="name">subjunctive</cell>
            <cell role="code">s</cell>
            <cell role="lang">ro</cell>
          </row>
          ...
        </table>
      </cell>
    </row>
    ...
  </table>
...
</div>
```

Fig. 1. Example of a MULTEXT-East common table: start of definition for Particle.



2.3.11. Particle

Table 13. Common specification for Particle

P	Attribute	Value	Code	English	Romanian	Russian	Czech	Slovene	Resian	Croatian	Serbian	Macedonian	Bulgarian	Estonian	Hungarian	Persian
0	CATEGORY	Particle	Q		ro		cs	sl	sl-rozaj	hr	sr	mk	bg			fa
1	Type	negative	z		ro					hr	sr		bg			
		infinitive	n		ro											
		subjunctive	s		ro											
		aspect	a		ro											
		future	f		ro											
		general	g										bg			
		comparative	c										bg			
		verbal	v										bg			
		interrogative	q							hr	sr		bg			
		modal	o							hr	sr		bg			
		affirmative	r							hr	sr					
2	Formation	simple	s									mk	bg			
		compound	c									mk	bg			
3	Clitic	no	n		ro											
		yes	y		ro											

Fig. 2. Example of a common tabels in HTML: Particle.

In MULTEXT and MULTEXT-East V3 the attribute-value definitions, together with MSD mapping information (i.e., the attribute position and the attribute-value code) were simply copied from the common tables. In MULTEXT-East V4 we take a more flexible position, where a language particular section can have a looser connection to the common tables — in fact, it could be a completely different specification, matching to the MULTEXT-East common one only in form. Of course, in this case any sensible mapping from the language particular specification to the common MULTEXT-East ones become very difficult, if not impossible. However, there do exist sensible compromises between the trivial mapping of MULTEXT and MULTEXT-East V3 and a completely unconstrained one.

The one we plan to adopt for the Slovene specification in Version 4 is exemplified by the JOS specification, where the tables will be aligned to the MULTEXT-East common ones in all respects, except for the attribute positions. This means that the feature-structure set of both will be identical, but not the MSDs. The reason for this is that MULTEXT-East has to cater for attributes of all languages, so language specific attributes (or those added to the specifications at a later date) wind up at the end of the string, leading to unwieldy MSDs, such as `Gppspe--n-----d`. This MSD has a number of hyphens only in order to maintain the position mapping to features, even though the attributes for some of these positions are never used for Slovene. With the freedom to reorder attributes, an individual language can use much shorter and more intuitive MSDs.

3 XSLT stylesheets

An important part of the specifications are the associated XSLT stylesheets, which allow for various transformations over the specifications. The stylesheets are written in XLST V1.0 and documented with XSLTdoc, *http://www.pnp-software.com/XSLTdoc/*. 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 transform or validate a list of MSDs.

```
<div type="section" xml:id="msd.N">
<head xml:lang="sl">Samostalnik</head>
<head xml:lang="en">Noun</head>
<table n="msd.cat" xml:id="msd.cat.N">
<head xml:lang="sl">Tabela atributov in vrednosti za samostalnik</head>
<head xml:lang="en">Attribute-Value Table for Noun</head>
<row role="type">
<cell role="position">0</cell>
<cell role="name" xml:lang="sl">samostalnik</cell>
<cell role="code" xml:lang="sl">S</cell>
<cell role="name" xml:lang="en">Noun</cell>
<cell role="code" xml:lang="en">N</cell>
</row>
<row role="attribute">
<cell role="position">1</cell>
<cell role="name" xml:lang="sl">vrsta</cell>
<cell role="name" xml:lang="en">Type</cell>
<cell role="values">
<table>
<row role="value">
<cell role="name" xml:lang="sl">občno_ime</cell>
<cell role="code" xml:lang="sl">o</cell>
<cell role="name" xml:lang="en">common</cell>
<cell role="code" xml:lang="en">c</cell>
</row>
<row role="value">
<cell role="name" xml:lang="sl">lastno_ime</cell>
<cell role="code" xml:lang="sl">l</cell>
<cell role="name" xml:lang="en">proper</cell>
<cell role="code" xml:lang="en">p</cell>
</row>
</table>
</cell>
</row>
```

Fig. 3. JOS morphosyntactic specifications: start of table for Noun.

3.1 Authoring

The two stylesheets belonging to this class are meant to assist in adding new languages to the specifications, and are the following:

msd-split.xsl makes a template for a language particular section on the basis of the value given to the `-langs` parameter, which should contain a space separated list of ISO language codes. So, to make section for a new language X, which is similar to Y and Z, the stylesheet would be run with -langs 'Y Z' and would produce a section with the union of the attribute-values for these two languages. These new language particular specifications are then corrected by hand.

msd-merge.xsl takes a language particular specification, and tries to "insert" it into the common specifications. This can mean simply adding the new language flags to existing attribute-value pairs, or adding new values or even new attributes to the common specifications.

3.2 Rendering

Displaying the stylesheets is currently only supported in HTML. This is done in two stages:

msd-spec2prn.xsl generates a "display-oriented" TEI document from the specifications. This means making display-oriented tables and generating the indexes of attributes, values, and MSDs.

msd-prn2html.xsl is a driver file, which calls the standard TEI stylesheets. It takes as input the display-oriented document and produces the HTML equivalent.

3.3 MSD conversion

The stylesheets in this class take a list of MSDs as a parameter, and, on the basis of the given specifications typically convert them to some form of feature-structures. The specifications can be either the MULTEXT-East common ones, or those for a particular language, depending on whether the MSDs are the common or language particular ones.

msd-expand.xsl produces different types of output, depending on the values of its "mode" parameter. It also takes parameters for input language (only MSDs valid for the language will be accepted) and for output language (it can localised to a language, which, of course, must be supported by the specifications). The output is in plain text tabular format, with columns that can be, depending on the value of mode, which is a space separated list of modes, the following:

- **check** only checks the validity of the input MSDs, flagging codes that are illegal for the language — this mode does not combine with the other ones;
- **id** identity transform (with possible localisation);
- **collate** collating sequence, with which it is possible to sort MSDs so that their order corresponds to the ordering of categories, attributes and their values in the specifications;
- **brief** expansion to values only, which the is the most compact feature-expanded format and is meant for short but still readable expansions of MSD; instead of binary values (yes/no), +/-Attribute is written;
- **verbose** expansion to feature-structures (attribute=value pairs) for all attributes defined for the category of the MSD;
- **canonical** expansion to feature-structures (attribute=value pairs) for all defined attributes, regardless of whether they are defined for a particular category or not;

msd-fslib.xsl transforms the MSD list into a XML/TEI feature and feature-structure libraries, suitable for inclusion into MSD annotated and TEI encoded corpora.

The intention isn't to run the above stylesheet whenever a transformation is needed but rather to run them, once the specifications are finished, over the complete set of MSDs to produce the tabular and XML files, which are then made available together with the specifications. To enable simpler processing and to produce output files with useful combinations of expansions, an additional Perl wrapper script is made available with the specifications.

4 Associated resources

Even though this paper is devoted to the morphosyntactic specifications, we also mention associated MULTEXT-East morphosyntactic resources, as without them, the specifications are not of much use. In the first instance this means the MULTEXT-East morphosyntactic lexicons, as it the lexicons that should provide the complete set of MSDs for a language, as well as examples of their usage. A second level resource are MSD annotated corpora, as this grounds the lexicon in contextualised examples of usage.

4.1 MULTEXT-East Lexicons

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 typically given as the equal sign; and (3) the *MSD*, i.e., the morphosyntactic description, which should be 1) valid according to the specifications and 2) contained in the set of MSDs listed in the lexical list of the language particular sections. It should be noted that this second criterion is to an extent circular, as it will be the lexicon that ultimately determines the list of valid MSDs; in practice, the process of constructing the MSD list and lexicon therefore typically proceeds in a cyclic fashion. Optionally, the lexicon can contain also contain (4) a column, giving the frequencies of the lexical entries in a corpus — for this, a MSD tagged and lemmatised corpus of the language must of course be available. Figure 4 gives some example entries from the Slovene lexicon.

```
alibi        =        Ncmsn
alibi        alibi    Ncmsa--n
alibija      alibi    Ncmda
alibija      alibi    Ncmdn
alibija      alibi    Ncmsg
alibije      alibi    Ncmpa
alibijem     alibi    Ncmpd
alibijem     alibi    Ncmsi
alibijema    alibi    Ncmdd
alibijema    alibi    Ncmdi
```

Fig. 4. Example of a MULTEXT-East morphosyntactic lexicons: the start of the paradigm for the Slovene masculine nominal lemma "alibi".

It is usually not the case that MULTEXT-East lexicons are produced from scratch but rather converted from some existing morphosyntactic lexica for a language. The MULTEXT-East lexica up to Version 3 were constructed according to different principles, but an ideal lexicon obeys the following principles:

1. The lexicon should contain all the valid MSDs for the language, even if only single exemplars are provided for particular MSDs. This criterion is in fact more strict than it seems, as languages with a large number of MSDs (e.g., Slovene has almost 2,000) exhibit a Zipfian distribution, i.e., quite a large number of MSDs can be quite rare in practice.
2. The lexicon should, for the lemmas it contains, include their complete inflectional paradigms. This is not always possible, as certain languages (e.g., agglutinating ones) can have "paradigms" with over a million word-forms but is manageable for even highly inflecting languages. The advantage is including the complete paradigms is that this makes the lexicon a very good resource for machine learning of lemmatisers; additionally, it also makes it more likely to obey the condition 1) above.
3. The lexicons should be of reasonable size (most current MULTEXT-East have around 15,000 lemmas), and, of course, the larger, the better. Ideally, the lemmas appearing in the lexicon should be grounded in an annotated corpus of the language, and the entries accompanied by corpus frequencies.

We do not here attempt to tackle the difficult problem of conversion of existing lexica to MULTEXT-East ones, but it should be noted that the `mtems-expand.xsl` in its `check` mode can be of considerable help in validating the lexical MSDs.

4.2 Annotated 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.

Corpora currently included in MULTEXT-East deliverables are all (translations of) the novel "1984" by G. Orwell. The complete novel has about 100.000 tokens, although this of course differs between the languages. The corpus is annotated with MSDs and lemmas, which makes it suitable for MSD tagging and lemmatisation experiments. Because it was the first such resource for many of the languages involved the annotation had to proceed mostly manually. The corpus is, in Version 3, encoded in XML, according to the Text Encoding Initiative Guidelines P4 [19], but it is planned to upgrade it to TEI P5 in Version 4. To exemplify the current structure, Figure 5 gives the start of the Slovene part of the corpus.

This parallel corpus also comes with separate alignment files, which contain, in V3, hand-validated pair-wise sentence alignments (not necessarily 1-1) between English and the translations. For V4 we also plan to provide pair-wise alignments between all the languages, which have been automatically induced from the alignments with English.

```
<text id="Osl." lang="sl">
 <body>
  <div type="part" id="Osl.1">
   <div type="chapter" id="Osl.1.2">
    <p id="Osl.1.2.2">
     <s id="Osl.1.2.2.1">
      <w lemma="biti" ana="Vcps-sma">Bil</w>
      <w lemma="biti" ana="Vcip3s--n">je</w>
      <w lemma="jasen" ana="Afpmsnn">jasen</w>
      <c>,</c>
      <w lemma="mrzel" ana="Afpmsnn">mrzel</w>
      <w lemma="aprilski" ana="Aopmsn">aprilski</w>
      <w lemma="dan" ana="Ncmsn">dan</w>
      ...
```

Fig. 5. Example of the annotation of the MULTEXT-East “1984” corpus: the start of the Slovene text “*Bil je jasen, mrzel aprilski dan*” (*It was a bright cold day in April*).

5 Conclusions

The paper presented the morphosyntactic specifications that will be part of the MULTEXT-East resources Version 4. The specifications currently cover 13 languages, and are encoded in TEI P5, with dedicated XSLT scripts to help with authoring the specifications for new languages, convert them into feature-structures or into a display HTML encoding. As the specifications cover a number of languages for which not many available and standardised resources exist, they can be a valuable reference point, and, together with the accompanying lexica and corpora, can serve as a “gold standard” dataset for language technology research and development, as well as for comparative linguistic studies.

There are a number of possible directions for further work. The language particular parts of the specifications could be further formalised and operationalised, esp. the combinations sections, as this would help in validating the MSD set for new languages. The attributes and their values could also be linked to other related attempts at standardisation of morphosyntactic features, in particular the ontology for descriptive linguistics GOLD *http://linguistics-ontology.org/gold.html* and the ISOcat Data Category Registry *http://www.isocat.org/*. There is also work to do in further formalisation of the MSDs and their relation to feature-structures, e.g., in allowing MSDs to include the metasymbols ’*’ or ’.’, i.e., having underspecified features in the MSD string.

Of course, we also hope that further languages will be added to the specifications. An obvious extension in this direction would be to add the original MULTEXT languages. However, we would encounter several problems: the specifications are incompatible outside the “common” features, so a way would needed to resolve this inconsistency, and in a backward compatible manner. More importantly, the associated resources, namely the lexicon and annotated corpus would have to be produced as well, to give the specifications some grounding in data. This is a relatively lengthily process, and it is unlikely that it could be carried out without dedicated international funding.

The situation is somewhat different, and better, for other, non Western European languages, where national efforts are underway to produce components of Basic Linguistic Resource Toolkits or BLARKs [16]; these can easily take the well-travelled route of developing MULTEXT-East compatible resources. Hopefully such an expansion could take place in the MONDILEX project, to include further Slavic languages into the specifications.

Finally, the most important aspect of the resources should be further encouraged, namely their use. Developing linguistic resources is not an end to itself, and they are worth only as much as they are used. We have therefore tried to maintain their quality and standardise their structure, to ensure that they can be interchanged and re-used for various purposes.

References

[1] Bel, N., Calzolari, N., and Monachini (eds.), M. (1995). Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets. MULTEXT Deliverable D1.6.1B, ILC, Pisa.

[2] Calzolari, N. and Monachini (eds.), M. (1996). Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora: A Common Proposal and Applications to European Languages. EAGLES Report EAG—CLWG—MORPHSYN/R, ILC, Pisa. http://www.ilc.cnr.it/EAGLES96/morphsyn/.

[3] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevič, V., and Tufiș, 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, Çanada. ACL.

[4] Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtsky, Z., and Žele, A. (2006). Towards a Slovene Dependency Treebank. In *Fifth International Conference on Language Resources and Evaluation, LREC'06*, Paris. ELRA.

[5] EAGLES (1996). Expert Advisory Group on Language Engineering Standards. http://www.ilc.pi.cnr.it/EAGLES/home.html.

[6] EAGLES (1996). Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG–TCWG–MAC/R, ILC, Pisa. http://www.ilc.cnr.it/EAGLES96/annotate/.

[7] Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC'04*, pages 1535 – 1538, Paris. ELRA. http://nl.ijs.si/et/Bib/LREC04/.

[8] Erjavec, T. (2006). MULTEXT-East Morphosyntactic Specifications and XML. In Slavcheva, M., Simov, K., and Angelova, G., editors, *Readings in multilinguality*, pages 41–48. Bulgarian Academy of Science, Sofia.

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

[10] Erjavec, T. and Krek, S. (2008). The JOS morphosyntactically tagged corpus of Slovene. In *Sixth International Conference on Language Resources and Evaluation, LREC'08*, Paris. ELRA.

[11] Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M., and Vitas, D. (2003). The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*, pages 25–32. ACL.

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

[13] Ide, N., Erjavec, T., and Tufiș, D. (2002). Sense Discrimination with Parallel Corpora. In *Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia. ACL.

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

[15] Ivanovska, A., Zdravkova, K., Erjavec, T., and Džeroski, S. (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.

[16] Maegaard, B., Krauwer, S., Choukri, K., and Jorgensen, L. D. (2006). The BLARK concept and BLARK for Arabic. In *Fifth International Conference on Language Resources and Evaluation, LREC'06*, Paris. ELRA.

[17] QasemiZadeh, B. and Rahimi, S. (2006). Persian in MULTEXT-East Framework. In *FinTAL 2006: 5th International Conference on Natural Language Processing*, pages 541–551, Turku, Finland.

[18] Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., and Divjak, D. (2008). Designing and evaluating a Russian tagset. In *Sixth International Conference on Language Resources and Evaluation, LREC'08*, Paris. ELRA.

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

[20] TEI Consortium, editor (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.*

[21] Tufiș, D. (1999). Tiered Tagging and Combined Language Model Classifiers. In Jelinek, F. and Noth, E., editors, *Text, Speech and Dialogue*, number 1692 in Lecture Notes in Artificial Intelligence, pages 28–33, Berlin. Springer-Verlag.

[22] Vojnovski, V., Džeroski, S., and Erjavec, T. (2005). Learning PoS Tagging from a Tagged Macedonian Text Corpus. In *Proceedings of the 8th International Conference Information Society, IS 2005*, Jožef Stefan Institute, Ljubljana.