[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

The MULTEXT-East Corpus

Tomaž Erjavec,[1] Nancy Ide[2]

(1) Dept. for Intelligent Systems
Institute Jožef Stefan,
Ljubljana

(2) Department of Computer Science
Vassar College
Poughkeepsie, New York

Abstract

The EU MULTEXT-East project has produced harmonised language resources for Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene. In this paper we introduce the MULTEXT-East multilingual corpus, which comprises marked-up texts in the six languages totaling approximately 2 million words and a small speech corpus. The corpus is encoded in SGML, in the TEI-like Corpus Encoding Specification and is divided into a parallel and a comparable (fiction, news) part. The parallel corpus consists of the novel "1984" by George Orwell in the English original and translations. The translations are sentence aligned with the original and tagged for word-level linguistic information, i.e. for morphosyntactic descriptions and lemmas. Detailed information on the corpus is available on the WWW and the corpus itself has been released for research purposes on a CD-ROM in the scope of the TELRI concerted action.

1. Overview

While standardised, large-scale language resources exist or are under development for most western languages there have, so far, been few comparable efforts for Central and Eastern European (CEE) languages. The MULTEXT-East (Multilingual Text Tools and Corpora for Eastern and Central European Languages) project (Erjavec et al., 1996) was a spin-off of the LRE project MULTEXT (Ide and Véronis, 1994) intended to fill this gap by developing language resources for six CEE languages (Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene) and adapting existing tools and standards to them. MULTEXT-East has extended MULTEXT's scope to CEE languages with the following goals:

- test and adaptation of language standards;
- development of an annotated multilingual corpus;
- development of morpho-lexical resources;
- adaptation of the MULTEXT corpus tools.

In this paper we concentrate on the MULTEXT-East multilingual corpus, which comprises marked-up texts in the six CEE languages totaling approximately 2 million words and a small speech corpus. The composition of the corpus is further explained in Section 2.

The MULTEXT-East corpus is encoded according to the Corpus Encoding Specification (CES) (Ide, 1998), suited for use in corpus linguistics and language engineering applications. The CES is based on and in broad agreement with the TEI Guidelines for Electronic Text Encoding and Interchange (Sperberg-McQueen & Burnard, 1994; see also Ide & Veronis, 1995). The CES identifies a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and linguistic information). The process of applying CES to new texts in new (MULTEXT-East) languages has led a major revision and extension of the CES. The CES encoding of the MULTEXT-East corpus is described in Section 3.

CES also provides encoding conventions for linguistic annotation, which is, however, not encoded in the corpus text itself (i.e. in the primary data) but as 'stand-off' annotation, that is, in separate documents hyperlinked to the primary corpus data. For the MULTEXT-East parallel corpus two such annotations were created. The translations of '1984' were aligned with the English original and the process of alignment and the resulting documents are discussed in Section 4. The '1984' was also annotated with word-level linguistic information, namely with context disambiguated morphosyntactic descriptions (Part-of-Speech tags) and lemmas; this is further explained in Section 5. Finally, Section 6 describes the current availability of the MULTEXT-East corpus.

2. Corpus composition

MULTEXT-East has built an annotated multilingual corpus, which is divided into a text corpus and a speech corpus. The text corpus consists of the parallel part and two comparable parts, where each of the three parts contains approximately 100.000 words per language.

The multilingual parallel corpus consists of the novel '1984' by George Orwell in the English original and translations into the six languages of the project. The choice of this data was motivated by the availability of the translations in all six languages and the availability of the English original and the Slovene translation in digital form from the Oxford Text Archive via the European Corpus Initiative. The ' 1984' is the central component of the MULTEXT-East corpus; while the whole corpus has been bibliographically

and structurally marked-up, the parallel part has been additionally sentence segmented and aligned as well as anno-
tated with word-level linguistic information.

The multilingual comparable corpus contains, for each of the six languages, a 'fiction' part and a 'news' part, where the data is comparable across the six languages in terms of the number and size of texts.

The fiction subset comprises either a single novel or ex-
cerpts from several novels:

- Bulgarian: one novel & four collections of short stories
- Czech: excerpts from the novel "Opera - pruvodce operni tvorbou" by Anna Hostomská
  - Estonian: 51 excerpts from novels
  - Hungarian: excerpts from four novels
- Romanian: two novelettes and a novel by Mihai Radulescu
- Slovene: the novel "Galjot" by Drago Jančar

The news subset comprises articles from daily newspapers:

- Bulgarian: "Kontinent" daily
- Czech: "Lidove noviny" daily
- Estonian: articles from 11 newspapers
- Hungarian: "Magyar Hirlap" daily
- Romanian: "Romania Libera" daily
- Slovene: "Dnevnik" daily

MULTEXT-East has also produced a small corpus of spoken texts. The text were taken from the EUROM-1 speech corpus and comprise the translations (from English) of forty short passages of five thematically connected sentences. These texts are CES encoded as the other parts of the corpus. Additionally, for Estonian, Hungarian, Romanian and Slovene, the texts have been spoken by a native speaker, these recordings digitised and stored in the EUROM format.

3. Primary Data Encoding

The entire text of the corpus[1] is encoded as a <**cesCorpus**> element, comprising the corpus header and the texts of the corpus, encoded as <**cesDoc**> elements. There are all together 26 texts, with 2 × 7 texts for the '1984' and 'speech' English-hub parallel parts and 2 × 6 for the two subsets of the comparable corpus.

The corpus header and each of the texts is stored in its own file and these files are given identifiers in the MULTEXT-East catalog, which is structured according to the SGML Open Technical Resolution 9401:1997. So, for example, the MULTEXT-East Estonian Fiction text has the formal PUBLIC identifier `-//MTE//TEXT CES1 Fiction//ET.`

For (language specific) character representation the documents use SGML defined entities from the entity sets `Added Latin 1`, `Added Latin 2`, `Russian Cyrillic`, and `Non Russian Cyrillic`. The last two are used for Bulgarian, the first two by all the other MULTEXT-East languages. Additionally, a few entities are used from `ISO 8879:19867/ENTITIES Publishing//EN.`

3.1. The Headers

The corpus as a whole as well as each of the <**cesDoc**> texts contains a header, which gives information about the corpus / text. The headers contain, much as in TEI, four elements:

- The <**fileDesc**> describes the corpus / text itself, where this description also contains the bibliographic information on its source.
- The encoding is detailed in the <**encodingDesc**> element where, e.g. the number of times particular elements are used in the corpus / texts is specified.
- The <**profileDesc**> gives the non-bibliographic information on the texts, in particular, it defines the languages that are used in the corpus and gives the™ I their identifiers. For language IDs the corpus header gives the definition of ISO 639 languages, e.g. the definition for the Slovene language is <**language id=sl iso639=sl**>Slovene</**language**> while the profile description of e.g. the Czech Orwell includes the element <**language id=ns-cs iso639=cs**>Newspeak Czech </**language**>.
- Finally, the <**revisionDesc**> gives the revision history of the corpus and its texts.

3.2. Structural markup

The texts of the corpus are marked up for gross structure (divisions, paragraphs, heads, lists, bylines, footnotes, etc.) and, depending on the text in question, various subparagraph markup, e.g. abbreviations, names, quotes, highlighted material, etc.

The parallel corpus was also automatically marked up for sentences (<**s**>) using the MULTEXT tools and **special** scripts and this markup hand validated. In inserting the <**s**> markup the well known problem of crossing <**q**> and <**s**> hierarchies appeared. This was solved by automatically splitting the <**q**> elements where necessary. The <**q**> elements that were so inserted were marked by **type=MI**, for "Machine Inserted".

All the structural elements of the corpus are also marked with IDs enabling direct reference to these elements. To exemplify the markup of the primary data we give below the beginning of the English Orwell text:

[1] The encoding of the primary data along with the format of the digital originals and the process that led from these to the CES encoding is given in (Erjavec (ed.), 1997).

```
<text>
 <body lang=en id=Oen>
  <div id="Oen.l" type=part n=l>
  <div id="Oen.l.l" type=chapter n=l>
  <p id="Oen.1.1.1">
  <s id="Oen.1.1.1.1">
  It was a bright cold day in April,
  and the clocks were striking
  thirteen.</s>
  <s id="Oen.1.1.1.2">
  <name type=person>Winston
  Smith</name>, his chin nuzzled into
  his breast in an effort to escape
  the vile wind, slipped quickly
  through the glass doors of <name
  type=place>Victory Mansions</name>,
  though not quickly enough to prevent
  a swirl of gritty dust from entering
  along with him.</s>
  </p>
```

4. Alignment

Each of the six translations of '1984' has been automatically S aligned with the English original and the alignments hand validated.[2] The process of validation was a cyclic one, with initial errors of alignment guiding both the harmonisation of gross document markup (paragraphs, lists, etc.) across languages and the correction of wrongly assigned sentence boundaries.

The alignments of '1984' are not hierarchical, i.e. division and paragraph level alignments have not been retained although they have been used in the process of alignment. The S-level elements that have been aligned are the following:

- <**s**> (sentence)
- <**item**> (an item in a <**list**>)
- <|> (a line in a <**poem**>)

Automatic alignment produced in addition to 1-1 links, 2-1, 1-2, 2-2, 0-1, and 1-0 links. In manual verification a number of other links were discovered as well. First, where there was a sequence of 0-1 or 1-0 links, these were (typically) merged into 0-n or n-0 links. Such links were due to translators not translating a portion of the text or, in certain rare cases (e.g. in poems) adding new text. Other, unexpected link types were also discovered, e.g. 1-6 and 2-4 links.

Each of the six pairwise alignments with English has been encoded as a separate SGML document with the <**cesAlign**> root element and stored in a separate file. The <**cesAlign**> documents do not contain the primary data but only <**link**>s to (S-level) elements, expressed as pairs of ID references to the parallel S-units of the two aligned <**cesDoc**> texts. The following hypothetical Slovene-English Orwell alignment span illustrates the syntax and types (one, many, zero) of the alignment links:

```
<link xtargets="Os1.l.l ; Oen.l.l">
<link xtargets="Osl.1.2 Osl.1.3 ; Oenl.l.2">
<link xtargets="Osl.1.4 ; ">
```

The first link encodes an 1-1 alignment, the second a 2-1 and the third an 1 -0 alignment.

5. Linguistic markup

In addition to the <**cesDoc**> encoding, the seven-language Orwell corpus is also available in a tokenised and morphosyntactically annotated form. Each text is encoded as a <**cesAna**> document, which contains ID references to the S-level elements of the <**cesDoc**>**,** as well as the tokenised and annotated primary data.

To arrive at this linguistically annotated encoding the following steps have been performed:

1. the <**cesDoc**> texts have been simplified and converted to <**cesAna**> encoding;
2. the texts were tokenised with the MULTEXT tools and the tokenisation hand validated;
3. the tokens were annotated with lexical, i.e. ambiguous morphosyntactic descriptions, lemmas and tags;
4. the lexical information was disambiguated.

The tokens in the <**cesAna**> documents are encoded in <**tok**> elements and are either words (which can, however, also be compounds or separable parts of words) or punctuation marks. The two are distinguished by the value of the token's **type** attribute. The values used are WORD for words and PUNCT for punctuation marks. The word or punctuation mark is contained in the <**orth**> element. The punctuation tokens are annotated with (unambiguous) corpus tags, which are uniform across the corpus. The following hypothetical example illustrates such <**cesAna**> markup:

```
<par from='Oen.1.1.1' >
  <s from='Oen.1.1.1.l'>
   <tok type=WORD><orth>It</orth></tok>
   <tok type=WORD><orth>was</orth></tok>
   <tok type=WORD><orth>a</orth></tok>
   <tok type=WORD><orth>bright</orth></tok>
   <tok type=WORD><orth>cold</orth></tok>
   <tok type=WORD><orth>day</orth></tok>
   <tok type=WORD><orth>in</orth></tok>
   <tok type=WORD><orth>April</orth></tok>
   <tok type=PUNCT><orth>,</orth>
     <ctag>COMMA</ctag></tok>
   <tok type=WORD><orth>and</orth></tok>
```

The word tokens are annotated both with ambiguous lexical information and with context-dependent, disambiguated information. The former is contained in the <**lex**> elements of the token, the latter in the <**disamb**> element(s). In general there may be more then one <**disamb**>

[2] A detailed discussion of the alignment process and encoding as well as of the word-level linguistic markup is given in (Priest-Dorman et al. (eds.), 1997).

element per one token, in cases where the tagger or human could not decide how to disambiguate the word.

Both <**lex**> and <**disamb**> elements contain the <**base**> (lemma) of the token, its morphosyntactic description <**msd**>,[3] and (for some of the languages) its corpus tag, <**ctag**>. The following example illustrates the annotation of the Slovene word 'je', here functioning as the non-negative indicative present tense copula verb in the third person singular, with the base form 'biti/to be':

```
<tok type=WORD>
  <orth>je</orth>
  <disamb>
    <base>biti</base>
    <msd>Vcip3s--n</msd>
  </disamb>
  <lex>
    <base>biti</base>
    <msd>Vcip3s--n</msd>
  </lex>
  <lex>
    <base>jesti</base>
    <msd>Vmip3s--n</msd>
  </lex>
  <lex>
    <base>on</base>
    <msd>Pp3fsg--y-n</msd>
  </lex>
</tok>
```

It should be noted that the correct annotation is given both in the <**disamb**> as well as in (one of) the <**lex**> elements. The morphosyntactic descriptions in the <**lex**> elements of a token thus correspond to what is commonly known in the tagger literature as the ambiguity class of the word.

As can be noticed, the <**cesAna**> documents produced in the project are maximal in terms of contained data (ambiguity classes) and annotation (annotations as elements, not attribute values, and no tag minimisation). As the documents are furthermore encoded with SGML entities, rather than the 8bit ISO character sets, the resulting files are rather large. However, the intention was to provide these resources for interchange and as self-contained as possible.

6. Availability

The complete documentation of the MULTEXT-East project together with HTML corpus 'samplers' is available on the WWW *(http://nl.ijs.si/ME/).* In the scope of the TELRI concerted action a two-volume CD-ROM has been released (Erjavec at al., 1998), entitled "East meets West: A Compendium of Multilingual Language Resources". This CD-ROM contains, on one volume the multilingual parallel Plato corpus developed in the scope of TELRI and on the second volume the results of the MULTEXT-East project, including the corpus discussed here. The corpus on the CD-ROM has been further enhanced with four new translations of '1984', namely Latvian, Lithuanian, Serbian, and Russian. These translations have been encoded in the same way as the MULTEXT-East corpus, i.e. as <**cesDoc**> documents, with the Latvian, Lithuanian, and Serbian translation sentence segmented and aligned with the English as well. The CD-ROM is currently being made available for research purposes only, on a per-cost basis.

[3] For a discussion of the MULTEXT-East morphosyntactic descriptions and lexical resources, see (Tufiş et al., 1998).

Acknowledgements

A number of people were involved in producing the MULTEXT-East corpus. In particular, the main persons responsible for the Bulgarian part of the corpus are L.Dimitrova and L.Sinapova; for Czech, V.Petkevič. J.Klímová, V.Schmiedtová, J.Hajič, and B.Hladká; for English, G.Priest-Dorman and the authors; for Estonian, H.J.Kaalep, V.Villandi, H.Orav, L.Paldre, U.Talvik and K.Muischnek; for Hungarian, C.Oravecz and L.Tihanyi; for Romanian, D.Tufiş, Ş.Bruda, C.Diaconu, L.Diaconu; and for Slovene, T.Erjavec, M.Romih, O.Vuković, A.Bizjak, and P.Jakopin. The work presented here was supported by the EU project COP 102, MULTEXT-East.

References

Erjavec,T. (ed.) (1997). Sample Corpus Collection and Preparation. MULTEXT-East Final Report D2.1F, Institute Jožef Stefan, Ljubljana.

Erjavec.T., Ide,N., Petkevič,V. & Véronis, J. (19%) Multext-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. Proceedings of the First European TELRI Seminar: Language Resources for Language Technology, 87-98.

Erjavec,T, Lawson.A., & Romary,L. (1998). East meets West: Producing Multilingual Resources in a European Context. This volume.

Ide.N. (1998). Corpus Encoding Standard: SGML guidelines for Encoding Linguistic Corpora. This volume, (see also *http://www.cs.vassar.edii/CES/)*

Ide,N. & J.Véronis. (1994). MULTEXT (Multilingual Tools and Corpora). Proceedings of the 14th International Conference on Computational Linguistics, COLING'94 (pp. 90-96). Kyoto, Japan. 1

Priest-Dorman.G., Erjavec.T., Ide.N., Petkevič.V. (eds.) (1997). Corpus Markup. MULTEXT-East Final Report D2.3F, Institute Jožef Stefan, Ljubljana.

Ridings,D. (1996). Text representation in PAROLE. Parole MLAP 63-386, Work package 4.1.3

Sperberg-McQueen,C.M. & Burnard,L. (eds.) (1994) Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford.

Tufiş.D., Ide.N. & Erjavec,T. (1998). Standardised Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages. This volume.