Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984

Tomaž Erjavec
Department of Intelligent Systems,
Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana, SLOVENIA
tomaz.erjavec@ijs.si

Abstract

The European MULTEXT-East project developed corpora, lexica and tools for English, Romanian, Slovene, Czech, Bulgarian, Estonian, and Hungarian. The centrepiece of the corpus is the novel "1984" in the English original and translations. The novel is sentence aligned and its words annotated for context disambiguated lemmas and morphosyntactic descriptions. The 7-way parallelism and the high-density of harmonised annotations make the corpus a unique dataset for studying word-class syntactic tagging, multilingual lexicon induction, sense disambiguation and other language engineering applications. Since their publication in 1998, the corpus and morpholexical resources of the project have been substantially revised and corrected. These resources have now been re-encoded in TEI and made freely available for research purposes on the WWW. The paper presents this new version of the annotated "1984" corpus.

1 Introduction

The MULTEXT-East project (Multilingual Text Tools and Corpora for Eastern and Central European Languages, '95–'97) was a spin-off of the EU MULTEXT project (Ide and Véronis, 1994) and developed language resources for six languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, as well as for English, as the 'hub' language of the project. It also adapted existing tools and standards to these languages. The main results (Dimitrova et al., 1998) were an annotated multilingual corpus (Erjavec and Ide, 1998) and lexical resources (Ide et al., 1998) for the seven languages.

The development of the morpholexical resources proceeded in three stages. First, morphosyntactic specifications were developed for the languages of the project. These define the so called morphosyntactic descriptions (MSDs), which express word-class syntactic information. The second stage was building word-form lexica, which cover the lexical stock of the corpus collected in the project. Finally, the developed lexica were used to MSD annotate the parallel portion of the MULTEXT-East corpus, namely the novel "1984" by G. Orwell, in the original and translations. While this corpus consists of only one novel of relatively modest size, it was the first morphosyntactically annotated corpus for most of the project languages. It has therefore served as a "gold standard" dataset for studying word-class syntactic tagging and similar applications. A further advantage of the resources is that they are harmonised: the morphosyntactic descriptions as well as the lexica and corpus text share a common specification across the seven languages.

One of the objectives of MULTEXT-East has been to make its resources freely available for research purposes. In the scope of the TELRI concerted action (Trans European Language Resources Infrastructure), the results MULTEXT-East have been released on CD-

ROM (Erjavec et al., 1998); since its release, the resources it contains have been used in a number of studies and experiments (Varadi, 1999; Tufiş, 1999; Hajič, 2000; Džeroski et al., 2000). In the course of such work, errors and inconsistencies were discovered in the MULTEXT-East specifications and data, which were subsequently corrected. But because this work was done at different sites and in different manners, the corpus encodings had begun to drift apart.

The EU project CONCEDE (Consortium for Central European Dictionary Encoding, '98–'00) comprised most of the same partners as MULTEXT-East, offered the possibility to bring the versions back on a common footing. The new release contains the revised and expanded morphosyntactic specifications, the revised lexica, and the significantly corrected and re-encoded "1984" corpus.

This paper details this new CONCEDE release of the MULTEXT-East language resources. Section 2 explains the overall encoding, structure and size of the corpus, Section 3 concentrates on corpus alignment, Section 4 deals with the linguistic annotation of the corpus, the morphosyntactic specification and associated lexica and Section 5 concludes by discussing the distribution of the corpus.

2 Corpus encoding and structure

The new release of the "1984" corpus is encoded in accordance with the recommendations of the Text Encoding Initiative, TEI P3 (Sperberg-McQueen and Burnard, 1994), using the `TEI.prose` base tag set and the following additional tags sets: `TEI.corpus`, which gives us the root element of the corpus and a more detailed structure of the corpus header; `TEI.linking` for pointer mechanisms; `TEI.analysis` for basic linguistic analysis (tokenisation); and `TEI.fs`, for feature structures, which encode our morphosyntactic descriptions and specification. Furthermore, we make use of ISO entity sets to descriptively encode the language specific characters in our corpus, e.g., `&ccaron;` for č.

The complete corpus is encoded as one SGML document, composed of the corpus

```
<text id="0en." lang="en">
 <body>
  <div type="part" id="0en.1">
   <div type="chapter" id="0en.1.1">
    <p id="0en.1.1.1">
     <s id="0en.1.1.1.1">
      <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>
      <w lemma="day" ana="Ncns">day</w>
      <w lemma="in" ana="Sp">in</w>
      <w lemma="April" ana="Ncns">April</w>
      <c>,</c>
      <w lemma="and" ana="Cc-n">and</w>
      <w lemma="the" ana="Dd">the</w>
      <w lemma="clock" ana="Ncnp">clocks</w>
      <w lemma="be" ana="Vais-p">were</w>
      <w lemma="strike" ana="Vmpp">striking</w>
      <w lemma="thirteen" ana="Mc">thirteen</w>
      <c>.</c>
     </s>
...
```

Figure 1: The TEI structure of the novel

TEI header and seven corpus elements, each, in turn, consisting of its text TEI header and the body of novel, either in English or one of the translations. The corpus and text headers supply information on the resource, e.g., the description of the file, together with its sources, the encoding description, giving details on e.g., normalisation procedures and tag usage, and the revision description.

The novel is composed of three parts plus the Appendix, "The principles of Newspeak" and each of these consists of a number of chapters, marked up using the ⟨**div**⟩ element with the appropriate **type** attribute. The divisions are then composed of paragraphs, and these of sentences.[1] All elements, down to the sentence level are given identifiers. Finally, the sentences contain words and punctuation marks, which can be qualified by their **type** and linguistic annotation. This structure is illustrated in Figure 1.

To give an impression of the size of the corpus, we give in Table 1 the sizes of the TEI files (without the headers, and 7bit encoded with entities) and the tag usage for the com-

[1] More precisely, these are paragraph-level and sentence-level elements, as also e.g., poems/lines map to ⟨**p**⟩/⟨**s**⟩.

Table 1: Sizes and tag usage in the "1984" corpus

	All	**En**	**Ro**	**Sl**	**Cs**	**Bg**	**Et**	**Hu**
Mb	31.7	4.0	4.7	4.0	4.3	7.2	3.5	4.0
⟨**div**⟩	200	29	28	29	29	29	27	29
⟨**p**⟩	9110	1287	1346	1288	1298	1322	1266	1303
⟨**s**⟩	46626	6737	6520	6689	6752	6682	6478	6768
⟨**c**⟩	125016	14138	16556	21486	20498	15153	19467	17718
⟨**w**⟩	618879	104286	101772	90792	79870	86020	75431	80708

```
<link xtargets="Osl.1.1 ; Oen.1.1">
<link xtargets="Osl.1.2 Osl.1.3 ; Oenl.1.2">
<link xtargets="Osl.1.4 ; ">
```

Figure 2: Example of stand-off bilingual alignment

plete corpus and separately for each language.

3 Alignment

Each of the six translations of "1984" has been automatically sentence aligned with the English original and the alignments hand validated. The alignments are encoded in separate documents containing references to sentence IDs, as specified by the `cesAlign` DTD, an application of the Corpus Encoding Standard (Ide, 1998). Figure 2 gives a hypothetical Slovene-English alignment span illustrating the syntax and types the alignment links: the first link encodes an 1-1 alignment, the second a 2-1 and the third an 1-0 alignment.

While bilingual alignments are certainly useful, full multilingual alignments are of even more value as here correspondences with languages other than English can be studied, for example between Slovene and Czech. In the CONCEDE release we therefore produced a new `cesAlign` document, which contains 7-way alignments between sentences. There is an obvious problem with computing such alignments, as there is no guarantee that the bi-alignments match up across the languages. We solved the problem in a simple manner, by retaining only those alignments that have the English sentence span identical in all the six bi-alignments.

In addition to the 7-way `cesAlign`, we also produced a 'knitted' version of this alignment. It is composed of translation units, each containing seven segments, and each of these the sentences of the aligned span; Figure 3 illustrates this format with the first 7-way aligned segment of the corpus.

4 Linguistic annotation

The greatest value of the "1984" corpus certainly comes from the fact that its words are linguistically annotated with context disambiguated lemmas (base-form of the word-form) and morphosyntactic descriptions. This makes the corpus a good dataset for experiments in automatic tagging and lemmatisation, as witnessed by experiments on Romanian (Tufiş, 1999), Hungarian (Varadi, 1999) and Slovene (Džeroski et al., 2000), or using the multilingual annotated corpus for testing various approaches to tagging (Hajič, 2000).

For linguistic annotation we use the default `TEI.analysis` attributes on ⟨**w**⟩, namely **lemma** and **ana**,[2] the latter defined as an `IDREF`, i.e., a reference to an identifier. For example, the Slovene word-form *novim* might be in the corpus annotated thus:

```
<w lemma="nov" ana="Afpmsi">novim</w>
```

The MULTEXT-East languages are inflectionally rich, which is reflected in the large number of distinct MSDs used in the corpus, and in the ratio of distinct word-forms to lemmas. To illustrate these quantities we give in

[2]This does not hold for Bulgarian, as it does not use MSDs in the corpus, but a reduced tagset; furthermore, unlike the other languages, the annotations were assigned automatically with little manual validation. To distinguish it from the other languages, Bulgarian uses, instead of **ana**, the `CDATA` **function** attribute.

```
<tu id="Ozz.1">
<seg lang="en"><s id="Oen.1.1.2.1"><w lemma="the" ana="Dd">The</w> <w lemma="hallway" ana="Ncns
<seg lang="ro"><s id="Oro.1.2.3.1"><w lemma="hol" ana="Ncmsry">Holul</w> <w lemma="bloc" ana="N
<seg lang="sl"><s id="Osl.1.2.3.1"><w lemma="ve&zcaron;a" ana="Ncfsn">Ve&zcaron;a</w> <w lemma=
<seg lang="cs"><s id="Ocs.1.1.2.1"><w lemma="chodba" ana="Ncfsn">Chodba</w> <w lemma="p&aacute;
<seg lang="bg"><s id="Obg.1.1.2.1"><w lemma="&vcy;" function="SP">&Vcy;</w> <w lemma="&kcy;&ocy
<seg lang="et"><s id="Oet.1.2.2.1"><w lemma="trepikoda" ana="Nc-sn">Trepikoda</w> <w lemma="hai
<seg lang="hu"><s id="Ohu.1.2.2.1"><w lemma="az" ana="Tf">Az</w> <w lemma="el&odblac;csarnok" a
</tu>
```

Figure 3: Example from the in-place 7-way alignment

Table 2 for the complete corpus and each language separately, the number of words in the corpus, the number of different word-forms (regardless of capitalisation or their annotation), and the number of different context disambiguated lemmas and MSDs.

4.1 Morphosyntactic specifications

The syntax and semantics of the morphosyntactic descriptions (MSDs) are given in the MULTEXT-East morphosyntactic specifications (Erjavec and (eds.), 1997). These specifications have been developed in the formalism and on the basis of specifications for six Western European languages of the EU MULTEXT project (Ide and Véronis, 1994) and in cooperation with EAGLES, the Expert Advisory Group on Language Engineering Standards. The MULTEXT-East morphosyntactic specifications contain, along with introductory matter, three parts: (1) the list of defined categories (parts-of-speech); (2) for each category a table of defined attribute-values (3) for each language, a language particular section

The common tables of the specification give, for each category, a table defining the attributes appropriate for the category, and the values defined for these attributes. They also define which attributes/values are appropriate for each of the MULTEXT-East languages. The structure of the tables facilitates the addition of new languages.

The MSDs are structured and more detailed than is commonly the case for part-of-speech tags; they are compact string representations of a simplified kind of feature structures. The first letter of a MSD encodes the part of speech, e.g., `Noun` or `Adjective`). The letters following the PoS give the values of the position determined attributes. The specifications define, for each part of speech, its appropriate attributes and their values and one letter codes. So, for example, the MSD `Ncmpi` expands to `PoS:Noun, Type:common, Gender:masculine, Number:plural, Case:instrumental`. It should be noted that in case a certain attribute is not appropriate (1) for a language, (2) for the particular combination of features, or (3) for the word in question, this is marked by a hyphen in the attribute's position. Slovene verbs in the indicative, for example, are not marked for gender or voice, hence the two hyphens in `Vcip3s--n`.

Technically, the specifications are a LaTeX document (with derived Postscript and HTML renderings), where the common tables are plain ASCII in a strictly defined format. For the purposes of the TEI encoding we have converted the common tables and the list of lexical MSDs to features and features structures respectively. In this encoding we followed Sperberg-McQueen and Burnard (1994, p.515ff), which gives an extended example involving a hypothetical markup of the British National Corpus with PoS tags.

First, we needed to define the list of all valid MSDs. This, of course, includes the MSDs used in the corpus, but also the MSDs culled from the lexicons. The MSDs are then encoded as a feature structure library, ⟨**fsLib**⟩, where each MSD is expressed as a feature structure specifying its **type** (the part of speech), the language(s) the MSD is appropriate for, and its decomposition into features, i.e., attribute/value pairs. Some exam-

Table 2: Inflection in the "1984" corpus

	All	**En**	**Ro**	**Sl**	**Cs**	**Bg**	**Et**	**Hu**
Words	618879	104286	101772	90792	79870	86020	75431	80708
Forms	106316	9181	14041	16401	17592	15093	16810	19180
Lemmas	56727	7059	7245	7903	9103	8510	8716	10043
MSDs	3218	134	393	1023	954	115	401	399

```
<fs type="Noun" select="cs"    id="Npmp"   feats="N1.p N2.m N3.p"></fs>
<fs type="Noun" select="bg ro" id="Npmp-n" feats="N1.p N2.m N3.p N5.n"></fs>
<fs type="Noun" select="bg"    id="Npmp-y" feats="N1.p N2.m N3.p N5.y"></fs>
<fs type="Noun" select="cs sl" id="Npmpa"  feats="N1.p N2.m N3.p N4.a"></fs>
<fs type="Noun" select="cs sl" id="Npmpd"  feats="N1.p N2.m N3.p N4.d"></fs>
```

Figure 4: Examples of MSDs as TEI feature structures

ples are given in Figure 4.

The morphosyntactic specifications define the attribute/value pairs referred to in the MSDs and are encoded as a TEI feature library, ⟨**fLib**⟩. For each feature we give, apart from its identifier, the languages it is appropriate for and the full name of its attribute, while its value is encoded as the content of the feature, as a symbol with the full name of its **value**. Examples of features referred to in Figure 4 are given in Figure 5.

4.2 Morphosyntactic lexica

In addition to the morphosyntactic specifications and the annotated corpus, the new release contains also the updated morphosyntactic lexica for the seven languages in the MULTEXT format (Ide et al., 1998). An entry in the such a lexicon consists of (1) the word-form, (2) the lemma, and (3) the MSD. The lexica contain, except for Estonian and Hungarian, where this is impossible due the agglutinating nature of the languages, full inflectional paradigms of each lemma, while at least all the lemmas appearing the corpus are contained in the lexica.

5 Conclusions

The paper introduced the second release of the MULTEXT-East "1984" morphosyntactically annotated corpus. We concentrated on the manner of its encoding and its quantitative aspects. The corpus is intended to serve as a "gold standard" dataset for studying word-class syntactic tagging and similar language engineering applications.

The first release of the MULTEXT-East resources had been published on the second volume of the "East Meets West" CD-ROM and distributed at cost by TELRI. While the CD-ROM has sold out, TELRI has recently made its contents available via its TRACTOR archive of computational tools and resources, at http://www.tractor.de/.

The new release of the MULTEXT-East resources is available directly from the Web, at http://nl.ijs.si/ME/V2/, where it can be downloaded after submitting an agreement limiting the use of the resources to research purposes.

Acknowledgements

The author would like to than the anonymous reviewers for their comments and suggestions.

A number of people were involved in producing the second version of the MULTEXT-East "1984" corpus: the Czech data was revised by V. Petkevič, Estonian by H.J. Kaalep, Hungarian by C. Oravecz and Romanian by D. Tufiş. In the first version, the novel in English had been tagged only automatically, and the MSDs were only incompletely disambiguated; for the second version, the complete novel was re-tagged by J. Szenthe, under the supervision of T. Varadi, and these annotations further elaborated by A.M. Barbu and D. Tufiş.

```
<f select="bg cs en et hu ro sl" id="N1.p" name="Type"><sym value="proper"></f>
<f select="bg cs en ro sl" id="N2.m" name="Gender"><sym value="masculine"></f>
<f select="bg cs en et hu ro sl" id="N3.p" name="Number"><sym value="plural"></f>
<f select="cs hu sl" id="N4.a" name="Case"><sym value="accusative"></f>
<f select="cs hu sl" id="N4.d" name="Case"><sym value="dative"></f>
<f select="bg ro" id="N5.n" name="Definiteness"><sym value="no"></f>
<f select="bg ro" id="N5.y" name="Definiteness"><sym value="yes"></f>
```

Figure 5: Examples of morphosyntactic specifications as TEI features

The work on the first version of the corpus was supported by the EU project COP 102, MULTEXT-East, and the work on the second version by the EU project PL96-1142 Concede and by EU concerted action TELRI-II. The work on the individual languages was further supported by various partners' grants and contracts.

References

Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimír Petkevič, and Dan Tufiş. 1998. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL '98*, pages 315–319, Montréal, Québec, Canada.

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

Tomaž Erjavec and Monica Monachini (eds.). 1997. Specifications and notation for lexicon encoding. MULTEXT-East Final Report D1.1F, Institute Jožef Stefan, Ljubljana, December.

Tomaž Erjavec and Nancy Ide. 1998. The MULTEXT-East corpus. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 971–974, Granada. ELRA.

Tomaž Erjavec, Ann Lawson, and Laurent Romary. 1998. East meets West: Producing Multilingual Resources in a European Context. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 233–240, Granada. ELRA.

Jan Hajič. 2000. Morphological Tagging: Data vs. Dictionaries. In *ANLP/NAACL 2000*, Seatle.

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.

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

Nancy Ide. 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/.

C. M. Sperberg-McQueen and Lou Burnard, editors. 1994. *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford.

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

Tamas Varadi. 1999. Morpho-syntactic ambiguity and tagset design for Hungarian. In *Proceedings of the EACL-99 Workshop on Linguistically Interpreted Corpora (LINC-99)*, Bergen. ACL.