Corpus Collection and Preparation

COP Project 106 MULTEXT-East

Work Package WP2 – Task 2.1
Deliverable D2.1 F

Final Report
17 December 1997

Workpackage Coordinator:
Tomaž Erjavec

Task Leaders:
LJUBLJANA - Tomaž Erjavec
AIX - Nancy Ide

Contributors:
Bulgarian: L.Dimitrova and L.Sinapova
Czech: V.Petkevič, J.Klímová and V.Schmiedtová
English: T.Erjavec and N.Ide
Estonian: H.J.Kaalep, V.Villandi, H.Orav, L.Paldre, U.Talvik and K.Muischnek
Hungarian: C.Oravecz and L.Tihanyi
Romanian: Ş.Bruda, C.Diaconu, L.Diaconu, and D.Tufiş
Slovene: T.Erjavec, M.Romih and O.Vuković

Contents

Chapter 1

Introduction

1.1 Background

This report documents the Deliverable D2.1 carried out within the framework of the Copernicus 106 MULTEXT-East (Multilingual Text Tools and Corpora for Eastern and Central European Languages) Project. It consisted of collecting and annotating a multilingual corpus for the six languages of the project, totaling cca 2 million words. The MULTEXT-East corpus has four components: a parallel corpus, two collections of comparable material, and a small parallel speech corpus.

In the scope of D2.1, the text collections were obtained, in most cases in pre-existing digital form and were up-translated to at least Level 1 (header & basic structure markup) of the Corpus Encoding Specification, while the speech corpus was translated, and, for most of the languages, recorded, digitised and encoded in accordance with EUROM recommendations. The parallel corpus has additionally been sentence segmented and strucural elements marked with unique identifiers.

In addition to basic structure encoding, the six translations of the parallel text corpus have been sentence aligned to the (English) original. Furthermore, the parallel corpus has been tokenised and annotated for parts-of-speech. For this additional annotation, stored in separate SGML documents, refer to MULTEXT-East Deliverable D23.

As modern day corpora go, the MULTEXT-East textual corpus is of modest size: each of the three sets for the six languages has approximately 100k words. While its small size limits the utility of such a corpus, it nevertheless represents a valuable resource: it is the first such collection for Central and East European Languages and for some of the partners, a first standardised corpus of their languages. At least the parallel part is also significantly annotated in a common scheme. Furthermore, the corpora are only a part of the MULTEXT-East deliverables, with e.g. lexicons covering the words in the corpus giving added value to the corpus itself.

In this report we document the acquisition and preparation of the data delivered as the MULTEXT-East corpus. The structure of the report is similar to that of the MLCC (Multilingual Corpora for Cooperation) report. For each corpus component, we begin with a general overview of the data, followed by a detailed description of (English), Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene corpora.

1.2 Description of the Corpus

The MULTEXT-East corpus is divided into a text corpus, and a speech corpus. The text corpus consists of the parallel part, as the basis for translation studies, and two parts to allow comparable studies to be carried out in different languages. The overall composition is thus as follows:

1. **Multilingual Parallel Corpus: 1984**
   This set (7 x 100k words) is the multilingual parallel corpus, consisting of the English original and the translated data in the six languages of the project. The parallel data chosen for the project was the novel "1984" by George Orwell. The choice of data was motivated by the availability of translations in all six languages and the availability of the English original in digital form from the Oxford Text Archive via the European Corpus Initiative. Details of this set can be found in Chapter 2.
2. **Multilingual Comparable Corpus: Fiction**
   The intention was to make this first part of the comparable corpus (6 x 100k words) a set of six complete novels written by native speakers of the MULTEXT-East languages. This goal was not met completely, due to copyright and obtainability problems. In some cases the language components of this part are therefore composed either from excerpts of novels or of collections of short stories. Details of this set can be found in Chapter 3.
3. **Multilingual Comparable Corpus: News**
   This second part of the comparable corpus (6 x 100k words) is composed of newspaper articles from the six countries of the project. Details of this set can be found in Chapter 4.
4. **Multilingual Speech Corpus: EUROM**
   Finally, the MULTEXT-East corpus comprises a small parallel speech corpus. For this corpus, a sample (200 sentences) of the English part of the EUROM1 multilingual speech database was selected. This text was translated into the six languages and, except for Bulgarian and Czech, recorded by one male speaker and digitised in accordance with the EUROM recommendations. Furthermore, the texts were also encoded in CES format, to make them part of the overall MULTEXT-East CES corpus. Details of this set can be found in Chapter 5.

For each corpus component that was obtained in digital form, the original file has also been preserved.

1.3 License agreements

Most of the partners have now obtained written license agreements for their component corpora. However, securing such agreements has been very time consuming and, in some cases, unsuccessful in the first instance. As there isn't much precedence in these matters in Central & Eastern European Countries, this is hardly surprising: ready made license agreements were not available, and publishers were in most cases wary of agreeing to have their data publicly available, fearing potential loss of royalties from e.g. third-part re-prints of their materials. The details of particular license agreements of the component corpora are given in the language sections of the next three chapters and in the headers of the respective corpus components.

Our intention is that the MULTEXT-East corpora be made publicly available. However, this is currently not possible for the full corpus, due to copyright considerations. As the partners are still pursuing copyright agreements, we are hopeful that the situation will change in the future.

1.4 Corpus Encoding

The text corpus, along with the textual part of the speech corpus has been encoded in SGML, in the Corpus Encoding Specification (CES) DTD. The CES DTD, along with documentation can be obtained from `http://www.cs.vassar.edu/CES/`.

The complete MULTEXT-East CES corpus is encoded as a **cesCorpus** element, which comprises a header and the 26 component parts (2 x 7 + 2 x 6) of the corpus, each encoded as a **cesDoc** element.

Each corpus **cesDoc** component is stored in a separate file. All system identifiers (i.e. filenames) are encapsulated in the MULTEXT-East catalog, which is structured according to the SGML Open Technical Resolution 9401:1997. There the corpus as a whole and its components are given the following PUBLIC identifiers:

```
-//MTE//DOCUMENT CES1//

-//MTE//TEXT CES1 1984//EN        -//MTE//TEXT CES1 Speech//EN
-//MTE//TEXT CES1 1984//BG        -//MTE//TEXT CES1 Speech//BG
-//MTE//TEXT CES1 1984//CS        -//MTE//TEXT CES1 Speech//CS
-//MTE//TEXT CES1 1984//ET        -//MTE//TEXT CES1 Speech//ET
-//MTE//TEXT CES1 1984//HU        -//MTE//TEXT CES1 Speech//HU
-//MTE//TEXT CES1 1984//RO        -//MTE//TEXT CES1 Speech//RO
-//MTE//TEXT CES1 1984//SL        -//MTE//TEXT CES1 Speech//SL

-//MTE//TEXT CES1 Fiction//BG     -//MTE//TEXT CES1 News//BG
-//MTE//TEXT CES1 Fiction//CS     -//MTE//TEXT CES1 News//CS
-//MTE//TEXT CES1 Fiction//ET     -//MTE//TEXT CES1 News//ET
-//MTE//TEXT CES1 Fiction//HU     -//MTE//TEXT CES1 News//HU
-//MTE//TEXT CES1 Fiction//RO     -//MTE//TEXT CES1 News//RO
-//MTE//TEXT CES1 Fiction//SL     -//MTE//TEXT CES1 News//SL
```

Each **cesDoc** element is also marked with the **lang** attribute that gives the language of the corpus component (but note that all **cesHeader** elements are marked as English). For language IDs, the two letter ISO 639 values have been used. The definitions for the CES **language** elements are encapsulated in the file with the PUBLIC identifier:

```
ISO 639-1988//ENTITIES Languages//EN
```

These entities encompass the European languages; we give here the definitions for the MULTEXT-East languages and languages referred to in the MULTEXT-East corpus:

```
<language id=en iso639=en>English</language>
<language id=bg iso639=bg>Bulgarian</language>
<language id=cs iso639=cs>Czech</language>
<language id=et iso639=et>Estonian</language>
<language id=hu iso639=hu>Hungarian</language>
<language id=ro iso639=ro>Romanian</language>
<language id=sl iso639=sl>Slovene</language>

<language id=de iso639=de>German</language>
<language id=fr iso639=fr>French</language>
<language id=ge iso639=ge>German</language>
<language id=it iso639=it>Italian</language>
<language id=la iso639=la>Latin</language>
<language id=lv iso639=lv>Latvian</language>
<language id=ru iso639=ru>Russian</language>
<language id=sp iso639=es>Spanish</language>
```

For (language specific) character representation, the documents use SGML entities from the following entity sets:

```
ISO 8879-1986//ENTITIES Added Latin 1//EN
ISO 8879-1986//ENTITIES Added Latin 2//EN
ISO 8879-1986//ENTITIES Russian Cyrillic//EN
ISO 8879-1986//ENTITIES Non Russian Cyrillic//EN
```

The last two are used for Bulgarian, the first two by all the other MULTEXT-East languages.

For each language (the reference to) its language specific character set entities has been encapsulated in files with the following PUBLIC identifiers:

```
-//MTE//ENTITIES Bulgarian//EN
-//MTE//ENTITIES Czech//EN
-//MTE//ENTITIES Estonian//EN
-//MTE//ENTITIES Hungarian//EN
-//MTE//ENTITIES Romanian//EN
-//MTE//ENTITIES Slovene//EN
```

It should be noted that, in the current version, only the complete `-//MTE//DOCUMENT CES1//` corpus constitutes a valid SGML document. To enable partial processing of the corpus components, a very simple mechanism has been adopted, which is likely to change in the future. Namely, each component text has a SGML PROLOG, but commented out. To take as an example the beginning of `-//MTE//TEXT CES1 1984//CS`:

```
<!--DOCTYPE cesDoc PUBLIC "-//CES//DTD cesDoc//EN" [
  <!ENTITY % ONECOMPONENT "INCLUDE">
  <!ENTITY ISOlang PUBLIC
           "ISO 639-1988//ENTITIES Languages//EN">
```

```
    <!ENTITY % MTEcs PUBLIC
             "-//MTE//ENTITIES Czech//EN">
    %MTEcs;
  ]-->
```

To process a single text, the comment markers (`--`) should be removed from this prolog. Note that in order to enable this kind of separate processing, marked sections have been made use of: if `ONECOMPONENT` is set to `INCLUDE`, then **language** definitions are included in the **cesDoc** header. If the corpus as a whole is processed, `ONECOMPONENT` is set to `IGNORE`, and the the **language** definitions are a part of the **cesCorpus** header.

The **cesDoc** corpus components have been encoded at least up to CES level 1. Level 1 markup includes a TEI-like header (file, encoding, profile and revision descriptions), and universal text elements down to the level of the paragraph, e.g. textual divisions, paragraphs, titles and headings, footnotes, tables and poems. Some CES 2 level markup has also been included, e.g. quoted material (quote, q, mentioned elements), rendition information, and, to varying degrees, abbreviations, dates, names, and numbers. The parallel corpus sets are furthermore marked up for sentences. Below we give a summary of the elements used, together with the number of times they appear in the corpus.

- **cesdoc** = 26; **text** = 26; **group** = 1; **body** = 28; **div** = 3309;
- **p** = 27993; **head** = 3262; **byline** = 1240; **caption** = 145; **closer** = 2; **dateline** = 220; **figdesc** = 115; **figure** = 115; **list** = 61; **note** = 52; **opener** = 290; **poem** = 90; **quote** = 756; **table** = 3
- **s** = 65758; **cell** = 75; **item** = 350; **l** = 456; **row** = 15
- **q** = 23947
- **author** = 168; **bibl** = 168; **corr** = 1; **docauthor** = 729; **ptr** = 39;
- **abbr** = 6059; **date** = 1835; **distinct** = 223; **foreign** = 919; **hi** = 5328; **label** = 2; **measure** = 18; **mentioned** = 1419; **name** = 41340; **num** = 5238; **ref** = 26; **sp** = 251; **speaker** = 6; **term** = 6; **time** = 87; **title** = 920

The details on the mark-up of the component corpora can be found in the 'Corpus Encoding' sections of the following chapters.

For the corpus components, HTML translations of **cesDoc** headers and samples of the texts are also available. These were obtained with custom-written CES 2 HTML tables, using the Fred software package (see `http://www.oclc.org/fred/`).

Chapter 2

Multilingual Parallel: Orwell's "1984"

2.1 Overview

The novel "1984" by George Orwell is the central component of the MULTEXT-East corpus: it is the parallel text, where the English original is sentence aligned with the six languages of the project, and each translation tagged for part-of-speech. Despite the small size of this parallel corpus (7 x 100k words), it can nevertheless constitute a valuable linguistic resource for the MULTEXT-East languages, especially as the project also delivers lexica which cover the word-forms of "1984".

It was therefore important to ensure that the CES markup of "1984" is similar across the languages, as much as allowed by the differences of translations. This process proceeded in a cyclic fashion, with initial errors of alignment guiding the harmonisation process. The greatest care was thus devoted to checking the structural markup (paragraphs, sentences, lines, items), while the density of sub-sentence markup differs more across the languages.

Below we give an estimate for the number of words, by language. The wordcounts were produced by removing the SGML tags from the texts and then using a 'wc'-like procedure.

English	104.302
Romanian	101.460
Slovene	91.619
Bulgarian	87.235
Czech	80.366
Hungarian	81.147
Estonian	79.334

The following sections give the details on the encoding of the "1984" corpus as a whole, which is followed by sections that detail the MULTEXT-East encoding of the English original, and the Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene translations. For each language we give a description of the corpus, its structure, the structure of the original from which the CES version was derived, and the markup process that led from one to the other.

Note that the added 'linguistic' markup of the "1984" corpus (i.e. alignment and part-of-speech annotation), which is contained in separate SGML documents, hyperlinked to the "1984" corpus, is described in the MULTEXT-East Deliverable D23.

2.2 Structure of the Corpus

To illustrate the kinds of elements used in the "1984" corpus, and to show their distributions across the language components we give the following table, which gives the contents of the <**tagUsage**> elements in the headers of the respective language components:

GI	**EN**	**BG**	**CS**	**ET**	**HU**	**RO**	**SL**
abbr	38	28	23	73	38	3	26
body	1	1	1	1	1	1	1
date	40	40	39	18	39		33
div	28	28	28	28	28	28	28
foreign	39	29	91	93	43	430	7
head	1	1	1	5	5	28	29
hi	103	103	75	183	71	413	242
item	4	4	4	4	4	4	4
l	32	26	33	32	32	26	34
list	1	1	1	1	1	1	1
mentioned	261	256	244	44	281		
name	1744	1704	2181	2457	1843	2157	1327
note	2	8	2	2	2	3	1
num	52	34	48	14	10		
p	1286	1321	1285	1289	1292	1335	1288
poem	10	7	11	10	10	7	10
ptr	2	8	1	2	2		1
q	2209	1203	2208	2192	2197	2137	2260
quote	35	34	36	35	35	23	35
s	6701	6649	6714	6658	6732	6487	6689
title	46	41	45	29	40	1	10
term			2				

Tag usage in Orwell's "1984"

For the purpose of alignment it was important to ensure that the gross structure of the 7 languages was as similar as possible. The translations were encoded taking the English digital version as as the norm, with errors of alignment guiding the harmonisation. The elements that have been harmonised can be seen in the table above, as they occur the same number of times in all the languages. In particular, each of the "1984"s has one <**body**>, which consists of three <**div type=part n=1, 2, 3**> and of one <**div type=part n=appendix**>. Each part, except the Appendix, is further subdivided into a number of <**div type=chapter n=1, 2, . . .** >.

2.2.1 The languages of "1984"

In addition to the 'standard' language IDs that the **lang** attribute can refer to, some elements in the "1984" corpus have been marked up with more specific language values. These are defined in the <**langUsage**> element in the <**cesDoc**> header of the particular language components. We give the definitions below:

```
<language id="ns" iso639="none">Newspeak</language>
<language id="ns-jg" iso639="none">Newspeak official jargon</language>
<language id="en-ck" iso639="none">British Cockney English</language>

<language id=bg-cl    iso639=bg>Bulgarian colloquial</language>
<language id=ns-bg    iso639=bg>Newspeak Bulgarian</language>
<language id=ns-jg-bg iso639=bg>Newspeak official jargon Bulgarian</language>

<language id=cs-cl    iso639=cs>Czech colloquial</language>
<language id=ns-cs    iso639=cs>Newspeak Czech</language>
<language id=ns-jg-cs iso639=cs>Newspeak official jargon Czech</language>

<language id=ns-et    iso639=none>Newspeak Estonian</language>

<language id=ns-hu    iso639=hu>Newspeak Hungarian</language>
<language id=ns-jg-hu iso639=hu>Newspeak official jargon Hungarian</language>

<Language id="ns-ro"  iso639="xx">Nouvorb&abreve;</language>

<language id=ns-sl    iso639=sl>Newspeak Slovene</language>
```

It can be noted that the above definitions still need to be harmonised. They also raise some interesting questions as to the definition of 'language'.

2.2.2 Sentence segmentation

Prior to sentence alignment, the "1984" had to be sentence segmented. This was done automatically, by using a combination of the `MtSeg` tool and special purpose scripts written by Greg Priest-Dorman of Vassar.

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** (Machine Inserted) in the translations, and with **type=broken**, with **prev** and **next** ID references in the English original. Where the <**hi**> elements spanned more than one <**s**>, these were corrected by hand.

The automatically inserted <**s**> elements were then hand validated via the process of alignment: where the S alignment between the English version and a translation was not one-to-one, there was a fair chance of it being not a difference in translation, but of a wrongly placed <**s**>. Such alignments were checked, and the <**s**> tags corrected where necessary.

2.2.3 ID marking

The `ider` program, written by Greg Priest-Dorman, was used to automatically assign unique identifiers to the following structural elements in the "1984" corpus:

- <**div**>, <**quote**>
- <**p**>, <**list**>, <**poem**>
- <**s**>, <**item**>, <**l**>
- <**q**>

The values of the **id** attribute are strings composed of the initial letter **O** identifying the **O**rwell corpus, the two letter ISO 936 code identifying the language component of the corpus (`en, bg cs, et, hu, ro, sl`), followed by numbers separated by periods; these give the position of the tag in the SGML tree of the document, rooted at the BODY element. The following example illustrates the ID naming scheme:

```
  <body lang="sl" id="Osl">
    <div id="Osl.1" type=part n=1>
    <head>Prvi del</head>
      <div id="Osl.1.2" type=chapter n=1>
        <head>I</head>
        <p id="Osl.1.2.2">
          <s id="Osl.1.2.2.1">
            Bil je jasen, mrzel aprilski dan in ure so bile trinajst.
          </s>
...
        <quote id="Osl.1.2.17" rend="CN IT"><p id="Osl.1.2.17.1">
            <s id="Osl.1.2.17.1.1">
              <date iso8601="1984-04-04">4. april 1984</date>
            </s>
          </p></quote>
...
        <quote id="Osl.1.8.18" rend="CN IT"><poem id="Osl.1.8.18.1">
            <l id="Osl.1.8.18.1.1">Tam pod kostanjevim drevesom</l>
            <l id="Osl.1.8.18.1.2">izdala si me,</l>
            <l id="Osl.1.8.18.1.3">izdal sem te,</l>
            <l id="Osl.1.8.18.1.4">ne da bi trenila z o&ccaron;esom.</l>
          </poem></quote>
```

2.3 English

Contributors: Nancy Ide, Greg Priest-Dorman (CNRS/Vassar), Tomaž Erjavec (IJS)

2.3.1 Description of the Corpus

The digital source of the English original of "1984" was obtained from the European Corpus Initiative, Multilingual Corpus 1 CD-ROM (as the sub-component `men08` of its `mul08` part), which also contains a Slovene, Croatian and Serbian translation. The source for the ECI-MC1 versions of "1984" was, in turn, the Oxford Text Archive, who already made an SGML encoding of the text, which, however, was not TEI-compliant.

Below we reproduce the content of the editorial file of the English "1984" on the ECI-MC1, `mul08/men08/eng18.edt`:

> For the OTA, all four versions were prepared by the OUCS KDEM service in 1985 for Dr David C Bennett of the School of Oriental And African Studies at London University. The OTA has copies of letters from all the publishers concerned which appear to indicate that academic use of the text is permitted. A.M. Heath (the agents for Orwell's estate) state that the purposes "must be solely for the use of research and not for sale in any way". We have been unable to make contact with Dr Bennett: the texts here have not been encoded or proofread in any way since they were produced (other than the English text, which was converted to an SGML like encoding by John Price-Wilkin, and subsequently automatically converted to conform to the OTA's dtd by myself and Alan Morrison.
>
> (ECI: LB, Nov 1992)

The above would seem to indicate that the OTA digital source for the English "1984" has the same distribution restrictions as those imposed by the MULTEXT-East project itself, namely, that the use of the materials is free for academic purposes. This, however, does not apply to the ECI edition of the corpus, which in its copyright statement explicitly forbids further distribution of its corpus.

The English MULTEXT-East Orwell has 91619 words.

2.3.2 Structure of the Corpus

The English "1984" corpus body consists of four <**div type=part**>. Each part (except <**div type=part n=APPENDIX**>) is further subdivided into a number of <**div type=chapter**>. In the English version, only the appendix is followed by a <**head**>.

The text is segmented into paragraphs, with the <**head**>, <**quote**>, <**note**>, and <**poem**> elements marked-up at the paragraph level.

Sub-paragraph tagging consists of <**hi**>, <**q**>, <**mentioned**>, and <**name**> and <**term**>. The tag <**mentioned**> is reserved for Newspeak words, while <**term**> gives the description of such Newspeak words. This is possibly an abuse of the CES, but CES currently does not give the <**gloss**> tag, which was used in the digital source for the descriptions of Newspeak words.

The first chapter is also additionally marked-up for <**date**>, <**foreign**>, and <**num**>.

Rendering information, where given, is a descriptive (e.g. `italics`) value of the **rend** attribute. Rendering has, except for capitalisation, been removed from the tag content.

The following is an example from the English "1984" corpus:

```
<p id="Oen.1.1.7">
<s id="Oen.1.1.7.1"><name type=org>Ministry of Truth</name>,
&mdash;
<name type=org lang=ns>Minitrue</name>,
in
<name type=language>Newspeak</name>
<ptr id="Oen.1.1.7.1.4" target="Oen.1.1.8" rend=asterisk>
&mdash; was startlingly different from any other object in sight.</s>
<s id="Oen.1.1.7.2">It
was an enormous pyramidal structure of glittering white concrete,
soaring up, terrace after terrace,
<num>300</num>
metres into the air.</s>
<s id="Oen.1.1.7.3">From where
<name type=person>Winston</name>
stood it was just possible to read, picked out on its white face in
elegant lettering, the three slogans of the
<name type=org>Party</name>:
<q id="Oen.1.1.7.3.3" rend="CE CA" type=slogan>War is peace</q>
<q id="Oen.1.1.7.3.4" rend="CE CA" type=slogan>Freedom is slavery</q>
<q id="Oen.1.1.7.3.5" rend="CE CA" type=slogan>Ignorance is strength.</q></s>
</p>
<note id="Oen.1.1.8" place=foot>
<name type=language>Newspeak</name>
was the official language of
<name type=place>Oceania</name>.
For an account of its structure and etymology see Appendix.
</note>
```

2.3.3 Structure of the Original

The ECI version which was used as the digital source was encoded in a TEI compliant DTD, but not proofread. For a history of it's evolution, see the introduction at the beginning of this section. The following is an example from the ECI `mul08/mul08a.eci`:

```
<p>The Ministry of Truth &mdash;  Minitrue, in Newspeak* &mdash;  was
startlingly different from any other object in sight. It was
an enormous pyramidal structure of glittering white
concrete, soaring up, terrace after terrace, 300 metres into
the air. From where Winston stood it was just possible to
read, picked out on its white face in elegant lettering, the
three slogans of the Party:
<q>WAR IS PEACE </q>
<q>FREEDOM IS SLAVERY </q>
<q>IGNORANCE IS STRENGTH</q>  The Ministry of Truth contained,
```

2.3.4 Markup Process

The ECI version which was used as the digital source was manually converted to CES-1 compliance. Original OCR mistakes were corrected with a spelling checker. The first chapter has been additionally marked-up.

2.4 Bulgarian

Contributors: Ludmila Dimitrova (IMI), Lydia Sinapova (IIT) and Kiril Simov (CLPP)

2.4.1 Description of the Corpus

The electronic version of the Bulgarian translation of "1984" was created manually from the printed edition, because the electonic version of the text doesn't exist. We do not possess a written permission from the publisher to use the text. As all other texts within the MULTEXT-EAST corpora, it can be used for research/academic purposes only. The book was published by the publishing house "Profizdat", Sofia, Bulgaria, in 1989 as the first and so far only edition of the book in question.

The Bulgarian version of "1984" has 87235 words. Microsoft Word 6.0 was used to count words. The text was spell-checked.

2.4.2 Structure of the Corpus

The Bulgarian "1984" corpus body consists of four <**div type=part**>. The <**div**> elements have the n attribute, giving the successive number of the appropriate level of the <**div**> (except for the appendix, where n=APPENDIX). The <**div**> elements hava also 'ID' attribute, whose value has the prefix '`Obg`' and the chapter numbers separated by periods, e.g. <**div type=chapter n=1 id="Obg.1.2"**>. Counting of chapters starts from 1 in every part. Each part (except <**div type=part n=APPENDIX**>) is further subdivided into a number of <**div type=chapter**>. In the Bulgarian version, only the appendix is followed by a <**head**>.

Further the text is segmented into paragraphs, with the <**quote**>, <**note**>, and <**poem**> elements marked-up at the paragraph level.

Sub-paragraph tagging consists of <**item**>, <**l**>, <**list**>, <**q**>, <**s**>, <**abbr**>, <**date**>, <**foreign**>, <**hi**>, <**mentioned**>, <**name**>, <**num**>, <**ptr**>, and<**title**>.

The text has been automatically sentence segmented, and the segmentation hand-validated.

The<**q**> tag is used to mark quoted dialogue. Some <**q**> elements have been split in this process; these are marked with type=MI, for "machine inserted". For the <**q**> tag, the attribute '`broken=yes`' is used when no sentence terminating punctuation appears between two dialogue fragments by the same speaker (either inside the <**q**> itself or in the intervening text between two <**q**> tags). The <**name**> tag is used for all proper nouns and noun phrases, denoting names.

Adjectives derived from proper nouns are not tagged.

All <**name**> tags contain the '`type`' attribute. The tag <**foreign**> is used only for those Newspeak words which are typographically distinguished in the printed version of the text, and for the Latin words.

The tags <**body**>, <**div**>,<**item**>,<**l**>,<**list**>,<**note**>,<**p**>,<**poem**>,<**q**>,<**quote**>, <**ptr**> and <**s**> have been automatically supplied with ”id” attribute with the prefix ”Obg” and subsequent numbers, separated by periods, showing their hierarchical position within the text..

All tags are used so that to be in harmony with the English 1984 for MULTEXT-East; the differences are due only to the differences between the English electronic text and Bulgarian printed text. Rendering information is included within the appropriate tags where necessary as a descriptive value of the ’`rend`’ attribute. Two rendition short-cuts are used:

- ’`rend=mdash`’ stands for ’`rend="PRE mdash POST mdash"`’
- ’`rend=dblq`’ stands for ’`rend="PRE ldquo POST rdquo"`’.

The value ’`rend="PRE mdash"` (or `"PRE ldquo"`) is used when the quoted dialogue ends up with the paragraph (there is no other typographical distinction). The value ’`rend="POST mdash"` (or `"POST rdquo"`) is used when there is no typographical distinction (except ordinary punctuation) for the beginning of the quoted dialogue. No default rendition is used.

The entire book is marked up using the same level of detail, i.e. no part is more detailed than the rest.

The following is an example from the Bulgarian “1984” corpus:

```
<p id="Obg.1.1.7">
<s id="Obg.1.1.7.1">
<name type=org>&Mcy;&icy;&ncy;&icy;&scy;&tcy;&iecy;&rcy;&scy;&tcy;&vcy;&ocy;&tcy;&ocy;
&ncy;&acy; &icy;&scy;&tcy;&icy;&ncy;&acy;&tcy;&acy;</name> &mdash;
<name type=org lang=ns-bg>&Mcy;&icy;&ncy;&icy;&pcy;&rcy;&acy;&vcy;</name>
&pcy;&ocy; &ncy;&ocy;&vcy;&gcy;&ocy;&vcy;&ocy;&rcy;
<ptr id="Obg.1.1.7.1.3" target="Obg.1.1.8" n=1>
&mdash; &rcy;&yacy;&zcy;&kcy;&ocy; &scy;&iecy;
&ocy;&tcy;&lcy;&icy;&chcy;&acy;&vcy;&acy;&shcy;&iecy; &ocy;&tcy;
&vcy;&scy;&icy;&chcy;&kcy;&ocy;
&pcy;&rcy;&iecy;&dcy; &ocy;&chcy;&icy;&tcy;&iecy; &mcy;&ucy;.</s>
<s id="Obg.1.1.7.2">&Tcy;&ocy;&vcy;&acy; &bcy;&iecy;
&ocy;&gcy;&rcy;&ocy;&mcy;&ncy;&acy;
&pcy;&icy;&rcy;&acy;&mcy;&icy;&dcy;&acy; &ocy;&tcy;
&bcy;&lcy;&iecy;&scy;&tcy;&yacy;&shchcy; &bcy;&yacy;&lcy;
&bcy;&iecy;&tcy;&ocy;&ncy;,
&kcy;&ocy;&yacy;&tcy;&ocy; &tcy;&iecy;&rcy;&acy;&scy;&acy;
&scy;&lcy;&iecy;&dcy; &tcy;&iecy;&rcy;&acy;&scy;&acy; &scy;&iecy;
&icy;&zcy;&dcy;&icy;&gcy;&acy;&shcy;&iecy; &ncy;&acy;
&tcy;&rcy;&icy;&scy;&tcy;&acy; &mcy;&iecy;&tcy;&rcy;&acy;
&vcy;&icy;&scy;&ocy;&chcy;&icy;&ncy;&acy;.</s>
<s id="Obg.1.1.7.3">&Ocy;&tcy; &mcy;&yacy;&scy;&tcy;&ocy;&tcy;&ocy; &scy;&icy;
<name type=person>&Ucy;&icy;&ncy;&scy;&tcy;&hardcy;&ncy;</name>
&mcy;&ocy;&zhcy;&iecy;&shcy;&iecy; &dcy;&acy;
&pcy;&rcy;&ocy;&chcy;&iecy;&tcy;&iecy;
&icy;&zcy;&pcy;&icy;&scy;&acy;&ncy;&icy; &scy;
&iecy;&lcy;&iecy;&gcy;&acy;&ncy;&tcy;&ncy;&icy;
&bcy;&ucy;&kcy;&vcy;&icy; &vcy;&hardcy;&rcy;&khcy;&ucy;
```

```
&bcy;&yacy;&lcy;&acy;&tcy;&acy; &fcy;&acy;&scy;&acy;&dcy;&acy;
&tcy;&rcy;&icy;&tcy;&iecy; &lcy;&ocy;&zcy;&ucy;&ncy;&gcy;&acy;
&ncy;&acy; &pcy;&acy;&rcy;&tcy;&icy;&yacy;&tcy;&acy;:
<q id="Obg.1.1.7.3.2" type=slogan rend="CE CA">&Vcy;&ocy;&jcy;&ncy;&acy;&tcy;&acy; &iecy; &mcy;&icy;&rc
<q id="Obg.1.1.7.3.3" type=slogan rend="CE CA">&Scy;&vcy;&ocy;&bcy;&ocy;&dcy;&acy;&tcy;&acy; &iecy; &rc
<q id="Obg.1.1.7.3.4" type=slogan rend="CE CA">&Ncy;&iecy;&vcy;&iecy;&zhcy;&iecy;&scy;&tcy;&vcy;&ocy;&t
</s></p>
```

2.4.3 Structure of the Original

There was no original electronic version. The book was typed in by Ludmila Dimitrova (IMI), Lydia Sinapova (IIT) and Kiril Simov (CLPP).

During the hand-validation process and comparison with the English version it was found out that some English sentences are not translated, and a paragraph had been moved to the next chapter. The original text has not been corrected.

2.4.4 Markup Process

The text of the Bulgarian "1984" was marked up during the typing process accordingly to CES1 conformance. The encountered typographical errors in the printed edition were corrected.

2.5 Czech

Contributors: Vladimír Petkevič (UTKL) and Jana Klímová (FFUK)

2.5.1 Description of the Corpus

The electronic version of the Czech translation of "1984" was obtained by OCR in the Institute of Theoretical and Computational Linguistics at the Faculty of Philosophy, Charles University, Prague.

The electronic version of the book can be used for research/academic purposes only. We do not possess a written permission from the publisher to use the text but there exists an oral agreement between the abovementioned publisher and us in this respect.

The book was published by the publishing house Naše vojsko, Prague, Czech Republic, in 1991 as the first and so far last edition of the book in question.

As computed by the Unix program `wc` over the whole CES-1 document, the Czech version of "1984" has 99180 words in 1231059 bytes.

2.5.2 Structure of the Corpus

The Czech "1984" corpus body consists of three **<div type=part>** and of one **<div type=appendix>**. Each part is further subdivided into a number of **<div type=chapter>**.

The <**div**> elements have the **n** attribute, giving the successive number of the appropriate level of the <**div**>, and the **id** attribute, whose value has the prefix **Ocs** and the chapter numbers separated by periods, e.g. <**div type=chapter n=2 id=Ocs.1.2**>.

The text within chapters is segmented into paragraphs. Detailed subparagraph markup was elaborated using the <**name**>, <**hi**>, <**num**>, <**q**>, <**abbr**>, <**foreign**>, <**date**> and other tags. The markup was performed semiautomatically. The text was spell-checked. In accordance with CES, only proper nouns have been tagged, while adjectives derived from proper nouns, e.g. *Winstonovo*, have not.

As the rendering information, italics, bold face and capitalisation was used. The entire book is marked up using the same level of detail, i.e. no part is more detailed than the rest. The following is an example from the Czech “1984” corpus:

```
<p id="Ocs.1.1.7">
<s id="Ocs.1.1.7.1"><name type=org>Ministerstvo pravdy</name>,
 v
<name type=language>newspeaku</name>
<ptr id="Ocs.1.1.7.1.3" target="Ocs.1.1.8" rend=asterisk>
<name type=org lang=ns-cs>Pramini</name>
se d&ecaron;siv&ecaron; li&scaron;ilo od v&scaron;ech ostatn&iacute;ch objekt&uring; v dohledu.</s>
<s id="Ocs.1.1.7.2">Byla to obrovsk&aacute; stavba tvaru pyramidy ze z&aacute;&rcaron;iv&ecaron;
b&iacute;l&eacute;ho betonu, kter&aacute; se terasovit&ecaron; vyp&iacute;nala do v&yacute;&scaron;ky
<num>300</num>
metr&uring;.</s> <s id="Ocs.1.1.7.3">Z m&iacute;sta, kde st&aacute;l
<name type=person>Winston</name>, se dala na b&iacute;l&eacute;m pr&uring;&ccaron;el&iacute;
p&rcaron;e&ccaron;&iacute;st ozdobn&yacute;m p&iacute;smem vyveden&aacute; t&rcaron;i hesla
<name type=org>Strany</name>:
<q id="Ocs.1.1.7.3.3" rend="CE CA" type=slogan>V&aacute;lka je m&iacute;r</q>
<q id="Ocs.1.1.7.3.4" rend="CE CA" type=slogan>Svoboda je otroctv&iacute;</q>
<q id="Ocs.1.1.7.3.5" rend="CE CA" type=slogan>Nev&ecaron;domost je s&iacute;la</q></s>
</p>
<note id="Ocs.1.1.8" place=foot>
<name type=language>Newspeak</name>
byl &uacute;&rcaron;edn&iacute;m jazykem
<name type=place>Oce&aacute;nie</name>. O jej&iacute; struktu&rcaron;e a etymologii viz Dodatek.
</note>
```

2.5.3 Markup Process

The scanned text of the Czech “1984” was spell-checked and marked up to CES1 conformance. In the process, a few typographical errors were encountered. These were corrected.

2.6 Estonian

Contributors: Heiki-Jaan Kaalep, Viire Villandi and Heili Orav

2.6.1 Description of the Corpus

There was no digital version available for the Estonian translation of “1984”, so the book had to be typed in. A written permission was acquired from the publisher and the translator of “1984” to freely distribute it in any form provided that no commercial benefit was sought, and to use it for research and academic purposes without restrictions.

The Estonian version of "1984" contains 79334 words, as indicated in the header of the tagged version.

2.6.2 Structure of the Corpus

The Estonian "1984" corpus body consists of three <**div type=part n=1, 2, 3**> and of one <**div type=part n=appendix**>. Each part is further subdivided into a number of <**div type=chapter n=1, 2, . . .** >. In the Estonian version, each <**div**> is followed by a <**head**>, giving the part or chapter number as it appears in the original text. Counting of chapters starts from **1** in every part.

All the tagging was done by hand.

The text is marked up to the level of sentences: <**p**>, <**quote**> plus marking of sub-paragraph element <**q**>, incl. broken Qs. Some marking of particular sub-paragraph elements.

<**body**>, <**div**>, <**head**>, <**item**>, <**l**>, <**list**>, <**note**>, <**p**>, <**poem**>, <**ptr**>, <**quote**>, <**text**> are used so that to be in harmony with the English electronic version of 1984 for MULTEXT-EAST v. 4; the differences are due only to the differences between the English electronic and Estonian printed version.

<**abbr**>, <**date**>, <**foreign**>, <**hi**>, <**mentioned**>, <**name**>, <**num**>, <**q**>, <**title**> are used sloppily.

Rendering information attribute has been in most cases included with the appropriated tags. The possible values for rend attribute are: **CA**, **IT**, , **smallpoint**, **PRE ldquor**, **POST ldquo**, **PRE ldquor POST ldquo**, **PRE rsquor POST rsquor**.

The following is an example from the Estonian"1984" corpus:

```
<p id="Oet.1.2.7">
<s id="Oet.1.2.7.1">
<name type=org>
T&otilde;eministeerium
</name>
 &mdash;
<name type=language>
uuskeeles
</name>
<ptr target=oet.N1 rend=asterisk>
<name type=org>
T&otilde;min
</name>
 &mdash;  erines rabavalt k&otilde;igest muust, mida oli n&auml;ha.
</s>
<s id="Oet.1.2.7.2">
See oli tohutu kiiskavvalgest betoonist p&uuml;ramiidne ehitis, mis kerkis
astanguliselt
<num>
300
</num>
```

```
  meetri k&otilde;rgusele.
</s>
<s id="Oet.1.2.7.3">
Sealt, kus
<name type=person>
Winston
</name>
  seisis, seletas silm veel parajasti valgel seinal elegantses kirjas
ilutsevat
<name type=org>
Partei
</name>
 kolme loosungit:
<q id="Oet.1.2.7.3.3" rend=CA type=slogan>
S&otilde;da on rahu
</q>
<q id="Oet.1.2.7.3.4" rend=CA type=slogan>
Vabadus on orjus
</q>
<q id="Oet.1.2.7.3.5" rend=CA type=slogan>
Teadmatus on j&otilde;ud
</q>
</s>
</p>
```

2.6.3 Structure of the Original

The original was typed in using a text editor under DOS. The Estonian characters were typed, using codes from extended ASCII. Some mark-up was included in the typing phase already. Follows an example from the original:

```
T=F4eministeerium - uuskeeles <note place=foot>Uuskeel oli
Okeaania ametlik keel. Selle struktuuri ja et=FCmoloogia kohta vt.
Lisa.</note> T=F4min - erines rabavalt k=F4igest muust, mida oli
n=E4ha. See oli tohutu kiiskavvalgest betoonist p=FCramiidne ehitis,
mis kerkis astanguliselt 300 meetri k=F4rgusele. Sealt, kus
Winston seisis, seletas silm veel parajasti valgel seinal
elegantses kirjas ilutsevat Partei kolme loosungit:

        SO^DA ON RAHU

        VABADUS ON ORJUS

        TEADMATUS ON JO^UD
```

2.6.4 Markup Process

The electronic version of "1984" was proofread and marked up to CES1 conformance. The whole Estonian "1984" CES1 corpus was cross-checked with the printed edition, and the printed edition was used to insert additional (e.g. <**hi**>) markup. In the process, a number of typographical errors were discovered in the electronic version of "1984".

2.7 Hungarian

Contributors: Csaba Oravecz and Laszló Tihanyi (RIL)

2.7.1 Description of the Corpus

Since there was no digital source serving as the basis of the CES1 encoded version of the Hungarian translation of "1984", the book had to be typed in. Copyright issues regarding the free use of the translation for academic and research purposes have been satisfactorily settled.

The Hungarian version contains 81167 words, as indicated in the header of the encoded corpus.

2.7.2 Structure of the Corpus

The Hungarian corpus of "1984" is composed of three <**div type=part**> and of one <**div type=appendix**>. Each part is further subdivided into a number of <**div type=chapter**>. In the Hungarian version, each <**div type=part**> is followed by a <**head**>, rendering the number of the part as it is indicated in literal written form in the printed edition.

The ¡div¿ elements have the n attribute, giving the successive number of the appropriate level of the ¡div¿, and the id attribute, the value of which is prefixed with Ohu, and contains a number scheme indicating the hierarchical position of the element in the sgml tree, e.g. ¡div id="Ohu.1.2" type=chapter n=1¿. The id attribute is also specified on each element down to the sentence level.

The text is segmented into paragraphs, with the ¡quote¿, ¡note¿, ¡poem¿ and ¡list¿ elements marked up at the paragraph level.

Sub-paragraph tagging is represented by ¡hi¿, ¡q¿ and ¡name¿. Frequently occurring names of people, places, organizations, products, languages, and events, are marked throughout the text.

Rendering information is specified in the header of the corpus file. If not explicitly indicated therein, it is included with the appropriate tag in the rend attribute. Possible values are: asterisk, IT for italics, CA for capitals, CE CA for centered caps, PRE mdash, PRE ldquor POST rdquor.

The following is an example from the Hungarian 1984 corpus:

```
<p id="Ohu.1.2.7">
<s id="Ohu.1.2.7.1">Az
<name type=org>Igazs&aacute;g-miniszt&eacute;rium</name> &mdash;
<name type=org lang=ns-hu>Minigaz</name>, ahogy
&uacute;jbesz&eacute;l&uuml;l<ptr id="Ohu.1.2.7.1.3"
target="Ohu.1.2.8" rend=asterisk> nevezt&eacute;k
&mdash; ijeszt&odblac;en el&uuml;t&ouml;tt a
k&ouml;rny&eacute;k&eacute;n
```

```
l&eacute;v&odblac; t&ouml;bbi &eacute;p&uuml;lett&odblac;l.</s>
<sid="Ohu.1.2.7.2">Ragyog&oacute;
feh&eacute;r betonb&oacute;l
k&eacute;sz&uuml;lt, &oacute;ri&aacute;si, piramis alak&uacute;
&eacute;p&iacute;tm&eacute;ny volt, s h&aacute;romsz&aacute;z
m&eacute;ter magasan ny&uacute;lt fel a leveg&odblac;be.</s>
<s id="Ohu.1.2.7.3">Onnan, ahol
<name type=person>Winston</name> &aacute;llt, &eacute;ppen el lehetett
olvasni a
<name type=org>P&aacute;rt</name>
h&aacute;rom jelmondat&aacute;t, amely d&iacute;szes
bet&udblac;kb&odblac;l volt kirakva az &eacute;p&uuml;let feh&eacute;r
homlokzat&aacute;ra:
<q id="Ohu.1.2.7.3.3"    rend="CE CA" type=slogan>
A h&aacute;bor&uacute;: b&eacute;ke
</q>
<q id="Ohu.1.2.7.3.4"    rend="CE CA" type=slogan>
A szabads&aacute;g: szolgas&aacute;g
</q>
<q id="Ohu.1.2.7.3.5"    rend="CE CA" type=slogan>
A tudatlans&aacute;g: er&odblac;
</q></s>
</p>
<note id="Ohu.1.2.8"    place=foot>Az &uacute;jbesz&eacute;l
<name type=place>&Oacute;ce&aacute;nia</name>
hivatalos nyelve. Nyelvtani rendszer&eacute;nek &eacute;s
sz&oacute;kincs&eacute;nek magyar&aacute;zat&aacute;t l&aacute;sd a
f&uuml;ggel&eacute;kben.</note>
```

2.7.3 Structure of the Original

The original was typed into Word for Windows 6.0, and then a number of conversion programs were made and used to convert it into ASCII. Rendition information was automatically extracted from the Word version and converted into mark-up, and then checked and supplemented by hand. Here follows an example from the DOS-text version:

```
<p>
Az Igazsg-minisztrium - Minigaz, ahogy jbeszll*
neveztk - ijeszten elttt a krnykn lv tbbi plet-

* Az jbeszl cenia hivatalos nyelve. Nyelvtani
rendszernek s szkincsnek magyarzatt lsd a
Fggelkben.
<pb=9>

tl. Ragyog fehr betonbl kszlt, risi, piramis alak
ptmny volt, s hromszz mter magasan nylt fel a
levegbe. Onnan, ahol Winston llt, ppen el lehetett
```

```
olvasni a Prt hrom jelmondatt, amely dszes betkbl volt
kirakva az plet fehr homlokzatra:
</p>

<p>
A HBOR: BKE
</p>
<p>
A SZABADSG: SZOLGASG
</p>
<p>
A TUDATLANSG: ER
</p>
```

2.7.4 Markup Process

The Word doc-file, converted into DOS text format, was the basis of the encoding. It was checked with the printed edition and corrections and additional markup were supplied by hand. A number of errors in the electronic version as well as in the printed edition have been detected. Corrections related to the latter are indicated in the header of "1984".

2.8 Romanian

Contributors: Dan Tufiş and Ştefan Bruda (RACAI), Lidia Diaconu, Călin Diaconu (ICI)

2.8.1 Description of the Corpus

The Romanian version of "1984" was typed in after the printed book published by the "Univers" Publishing house, with a translation due to Mihnea Gafiţţa. The copyright problems are still questionable. The "Univers" Publishing house had a limited copyright which expired at the beginning of this year. The type-in process introduced a lot of errors. The proofreading, done by a different person than the typist elliminated much of the errors. During the dictionary construction several other errors showed up and were corrected.

As computed by the Unix program `wc` over the whole CES-1 document the Romanian version of "1984" has 104580 words.

2.8.2 Structure of the Corpus

The Romanian "1984" corpus body consists of three <**div type=part**> and of one <**div type=appendix**>. Each part is further subdivided into a number of <**div type=chapter**>. In the Romanian version, each <**div**> is followed by a <**head**>, giving the part or chapter number.

The <**div**> elements have the **n** attribute, giving the successive number of the appropriate level of the <**div**>, and the **id** attribute, whose value has the prefix **ro1984** and the chapter and section numbers separated by periods, e.g. <**div type=part n=1 id=ro1984.1.1**>.

The text is segmented into paragraphs, with the <**head**>, <**quote**>, <**note**>, and <**poem**> elements marked-up at the paragraph level.

Sub-paragraph tagging consists of <**hi**>, <**q**>, <**foreign**>.

Rendering information, given as the CES conformant two-letter value of the **rend** attribute has been in most cases included with the appropriated tags, except for the default preceding mdash of the <**q**> tag.

In our markup, the mdash had beeen replaced by simple quotes i.e. **rend=”PRE lsquo POST rsquo”** (preceded by left side quote, followed by right side quote).

The tag usage for the “1984” corpus is shown below.

```
<tagsdecl>
   <tagusage gi=body occurs=1></tagusage>
   <tagusage gi=div occurs=28></tagusage>
   <tagusage gi=head occurs=28></tagusage>
   <tagusage gi=hi occurs=410></tagusage>
   <tagusage gi=l occurs=30></tagusage>
   <tagusage gi=note occurs=3></tagusage>
   <tagusage gi=p occurs=1278></tagusage>
   <tagusage gi=poem occurs=6></tagusage>
   <tagusage gi=q occurs=996></tagusage>
   <tagusage gi=quote occurs=186></tagusage>
   <tagusage gi=text occurs=1></tagusage>
   <tagusage gi=foreign occurs=421></tagusage>
 </tagsdecl>
```

The following is an example from the Romanian “1984” corpus:

```
<p id="Oro.1.2.8">
<s id="Oro.1.2.8.1"><name type="org">Ministerul Adev&abreve;rului</name> -
&icirc;n <foreign lang="ns-ro">Nouvorb&abreve;</foreign>,
<foreign lang="ns-ro">Minadev</foreign> - te izbea fiindc&abreve; era
cu totul diferit de oricare alt&abreve; cl&abreve;dire care se
vedea.</s> <s id="Oro.1.2.8.2">Era o structur&abreve;
&icirc;nalt&abreve;, imens&abreve;, de beton armat alb,
str&abreve;lucitor, &icirc;n form&abreve; de piramid&abreve;, care se
ridica, teras&abreve; dup&abreve; teras&abreve;, la trei sute de metri
de la p&abreve;m&acirc;nt.</s>
<s id="Oro.1.2.8.3">De unde st&abreve;tea <name
type="person">Winston</name> se puteau citi foarte bine cele trei
lozinci ale <name type="org">Partidului</name>, scrise cu litere
elegante pe fa&tcedil;ada cea alb&abreve;:
<q id="Oro.1.2.8.3.3" rend="CA CN" type="slogan">R&Abreve;ZBOIUL ESTE PACE</q>
<q id="Oro.1.2.8.3.4" rend="CA CN" type="slogan">LIBERTATEA ESTE SCLAVIE</q>
<q id="Oro.1.2.8.3.5" rend="CA CN" type="slogan">IGNORAN&Tcedil;A ESTE
PUTERE</q></s>
</p>
```

2.8.3 Structure of the Original

There was no original electronic version. The book was typed-in. Apparently, either the Romanian Publisher or the translator took some liberty in defining paragraphs and this is why the paragraphs in the Romanian version do not match the paragraphs in the English version (ECI edition). However, this is does not happen many times.

2.8.4 Markup Process

The whole Romanian "1984" CES1 corpus was cross-checked with the printed edition, and the printed edition was used to insert additional (e.g. <**hi**>) markup. In marking the paragraphs we followed the Romanian published version.

2.9 Slovene

Contributors: Tomaž Erjavec (IJS) and Olga Vuković

2.9.1 Description of the Corpus

The digital source of the Slovene version of "1984" was obtained from the European Corpus Initiative, Multilingual Corpus 1 CD-ROM, which also contains the English original, as well a Croatian and a Serbian translation. The source for the ECI-MC1 versions of "1984" was, in turn, the Oxford Text Archive. For the OTA, all four versions were prepared by the OUCS KDEM service in 1985 for Dr David C Bennett of the School of Oriental And African Studies at London University. The OTA has copies of letters from all the publishers concerned which appear to indicate that academic use of the text is permitted.

The above would seem to indicate that the digital source for the MULTEXT-East Slovene "1984" has the same distribution restrictions as those imposed by the MULTEXT-East project itself, namely, that the use of the materials is free for academic purposes.

It is not clear what the Slovene printed version serving as the source of OTA's digital version was; however, it seems that the only translation and publication of "1984" in Slovene is the (out of print) one by "Mladinska knjiga", Ljubljana, published in 1983. Typical errors in the OTA edition further indicate that the OTA edition was obtained by OCR from the printed edition.

We currently do not have written permission to use the text, however, we do have verbal assurance from OTA as well as by employees of MK that it is acceptable to use the Slovene translation of "1984" for the purposes of the MULTEXT-East project.

The Slovene version of "1984" contains 91.619 words, as indicated in the header of the tagged version.

2.9.2 Structure of the Corpus

The Slovene "1984" corpus body consists of three <**div type=part n=1, 2, 3**> and of one <**div type=part n=appendix**>. Each part is further subdivided into a number of <**div type=chapter n=1, 2, . . .**>. In the Slovene version, each <**div**> is followed by a <**head**>, giving the part or chapter number. Counting of chapters starts from **1** in every part.

Elements <**body**>, <**div**>, <**head**>, <**item**>, <**l**>, <**list**>, <**note**>, <**p**>, <**poem**>, <**ptr**>, <**quote**>, <**text**> are used so that to be in harmony with the English 1984 for MULTEXT-

East; the differences are due only to the differences between the English electronic and Estonian printed version.

Sub-paragraph tagging, i.e. <**abbr**>, <**date**>, <**foreign**>, <**hi**>, <**mentioned**>, <**name**>, <**num**>, <**q**>, <**title**> are used sloppily.

For example, names have been tagged automatically, and the tagged names hand-validated; therefore all the tagged names are correct, but the text might also contain untagged names. The name tags do not contain the type attribute. In accordance with CES, only proper nouns have been tagged, while adjectives derived from proper nouns, e.g. *Winstonov*, have not.

Rendering information is given as the CES conformant two-letter value of the **rend** attribute. It has been in most cases included with the appropriated tags, except for the default preceding `—` of the <**q**> tag. The values for **rend** attribute are: CA, CE, CN, IT, and '*'.

Rendering, including 'all caps' capitalisation, has been removed from the tag content.

The first chapter has more detailed mark-up than the rest of the novel, e.g. it includes more **type** information on tags.

The text has been automatically sentence segmented, and the segmentation hand-validated. Some <**q**> elements have been split in this process; these are marked with **type=MI**, for 'machine inserted. The <**body**>, <**div**>, <**quote**>, <**p**>, <**poem**>, <**list**>, <**l**>, <**item**>, <**s**>, and <**q**> tags have been marked with the **id** attribute.

The following is an example from the Slovene "1984" corpus:

```
<p>
</p><p id="Osl.1.2.8">
<s id="Osl.1.2.8.1">Ministrstvo resnice &mdash; <name>Minires</name> v
<name>Novoreku</name><ptr id="Osl.1.2.8.1.3" n="1" rend="*" target="Osl.1.2.9">
&mdash; se je osupljivo lo&ccaron;ilo od kateregakoli predmeta.</s>
<s id="Osl.1.2.8.2">Bilo je velikanska
piramidasta zgradba iz ble&scaron;&ccaron;e&ccaron;e belega betona, ki
je v terasah kipela kvi&scaron;ku, tristo metrov visoko v zrak.</s>
<s id="Osl.1.2.8.3">S
kraja, kjer je stal <name>Winston</name>, se je ravno &scaron;e dalo
prebrati tri partijske parole, ki so se v lepih &ccaron;rkah
odra&zcaron;ale z belega pro&ccaron;elja:
<q id="Osl.1.2.8.3.2" rend="CN CA" type=slogan>Vojna je mir</q>
<q id="Osl.1.2.8.3.3" rend="CN CA" type=slogan>Svoboda je su&zcaron;enjstvo</q>
<q id="Osl.1.2.8.3.4" rend="CN CA" type=slogan>Nevednost je mo&ccaron;</q></s>
</p>
```

2.9.3 Structure of the Original

The ECI version which was used as the digital source was essentially the same as the OTA one, except that <**p**> had been automatically inserted into the text. Also, neither edition had been proofread or checked in any way. The Slovene characters were encoded as cy, sy, zy for č, š, ž respectively, which caused problems for Syme. The following is an example from the ECI version:

```
pojavljale brez ozadja in so bile vecyinoma nerazumljive.  Ministrstvo
resnice -- Minires v Novoreku* -- se je osupljivo locyilo od
```

```
kateregakoli predmeta. Bilo je velikanska piramidasta zgradba iz
blesycyecye belega betona, ki je v terasah kipela kvisyku, tristo
metrov visoko v zrak. S kraj a, kjerje stal Winston, seje ravno sye
dalo prebrati tri partijske parole, ki so se v lepih cyrkah odrazyale
z belega procyelja:
                             VOJNA JE MIR
                        SVOBODA JE SUZYNJSTVO
                           NEVEDNOST JE MOCY

</p>
```

2.9.4 Markup Process

The ECI edition was taken as the basis for the encoding. It was proofread and marked up to CES1 conformance. The whole Slovene "1984" CES1 corpus was cross-checked with the printed edition, and the printed edition was used to insert additional (e.g. <**hi**>) markup. In the process, a number of typographical errors were discovered not only in the ECI edition, but also in the printed edition of the Slovene translation of "1984".

Chapter 3

Multilingual Comparable 1: Fiction

This chapter gives the details of the 'Fiction' part of the MULTEXT-East multilingual comparable corpus. The intention was to make this set from six complete novels written by native speakers of the respective languages of the project. This goal was not met completely, due to problems of copyright and obtainability. The composition of this set is therefore as follows:

MULTEXT-East Multilingual Comparable Corpus, Fiction:

- Bulgarian: one novel & four collections of short stories
- Czech: excerpts from the novel: Anna Hostomská "Opera – pruvodce operní tvorbou"
- 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 following sections give the details for each of the above language components in turn.

3.1 Bulgarian

Contributors: Institute of Mathematics, Bulgarian Academy of Sciences

3.1.1 Description of the Corpus

The Bulgarian MULTEXT-East fiction corpus contains the novel "PASSION or the death of Alice" by Emilia Dvorianova and and the first four chapters of the novel "I want, I believe, I can" by Julia Berberyan.

The first novel was first published by "OBSIDIAN", Sofia in 1995, the second - by "Abagar Holding", Sofia in 1995. The Bulgarian site obtained the author's / publisher's agreement allowing the use of novels for the purposes of the MULTEXT-East project.

“PASSION or the death of Alice” was provided in WINWORD format, while “I want, I believe, I can” in ’Page maker’ format. The texts were converted to ASCII files.

The Bulgarian Fiction corpus contains 98371 words. The first part (“PASSION or the death of Alice”) contains 54100 words approx., the second part - 44200 words approx.

3.1.2 Structure of the Corpus

The corpus body consists of 15 <**div type=chapter**>, the number of each chapter is given in the attribute n of the tag. The chapters from “I want, I believe, I can”start with a <**head**> giving the title of the chapter.

The <**div type=chapter**> elements have the **n** attribute, giving the chapter number, and the **id** attribute, whose value has the prefix **Fictbg** followed by a period, a string, identifying the book, and the chapter number, for example <**div type=chapter n=1 id=Fictbg.alice.1.1**>.

The text is segmented into paragraphs, with the <**head**>, <**quote**>, <**opener**>, <**poem**> elements marked-up at the paragraph level.

Sub-paragraph tagging consists of <**abbr**>, <**foreign**>, <**hi**> <**date**>, <**name**>, <**l**>, <**num**> and <**q**>. Direct speech has been marked-up by <**q**>. “PASSION or the death of Alice” is tagged in full at the sub-paragraph level. “I want, I believe, I can” is tagged partially.

Rendering information, given as the CES conformant two-letter value of the **rend** attribute has been included with the appropriated tags and, for mdash and capitalisation, retained in the tag content.

The following is an example from the Bulgarian Fiction corpus; the characters are, in the original, encoded as SGML ISO Cyrillic entities (e.g. `о`); to make the following sample more readable, they were automatically transliterated into Latin (e.g. `о` ⇒ `o`).

```
<p>
Poslie vlyazokh i mie porhardsi dhardzhd ot shubraka, i pak sie
usietikh razshardrdiena, dazhie promhardrmorikh nieshcho nie
shardvsiem za kazvanie, shchoto i za tozi shubrak smie sie
razpravyali, no za niego gospozhitsata bieshie nieprieklonna i nie sie
usmikhvashie, a ustnichkitie ig sie svivakha na dhardgichka:
</p>
<p>
<q rend="PRE mdash">
Da nie si go
pipnala, lielsofto <name
type=person>Jo</name>!</q>
</p>
```

3.1.3 Structure of the Original

The versions which were the basis for the encoding, were ASCII files. The text contained a number of misrendered punctuation and spelling mistakes. A spell-checker was used to correct the spelling mistakes. It should be noted that the original electronic texts contained Latin letters used as cyrillic (e.g. latin “e” instead of cyrillic “e”) Their occurrences were discovered by the spell-checker, but we had to substitute latin with cyrillic letters manually. In the following example, the Cyrillic letters have been transliterated back into Latin:

```
Poslie vlyazokh i mie porhardsi dhardzhd ot shubraka, i pak sie
usietikh razshardrdiena, dazhie promhardrmorikh nieshcho nie
shardvsiem za kazvanie, shchoto i za tozi shubrak smie sie
razpravyali, no za niego gospozhitsata bieshie nieprieklonna i nie sie
usmikhvashie, a ustnichkitie ig sie svivakha na dhardgichka: - Da nie
si go pipnala, lielsofto Jo!
```

3.1.4 Markup Process

The version has been marked up to CES1 conformance manually - all tags were inserted manually.

3.2 Czech

Contributors: Vladimír Petkevič (FFUK) and Věra Schmiedtová (ÚJČ AV ČCR)

3.2.1 Description of the Corpus

The Czech MULTEXT-East fiction corpus is represented by a book entitled OPERA — pruvodce operní tvorbou (Opera — the Guide through the Art of Opera). The book written by Anna Hostomská was first published in 1955 by SNKLU, Prague, its second edition (Praha, Svoboda 1993) was the source of the Czech fiction corpus. The publication contains a lot of descriptions of the content of operas and descriptions of their authors. Thus it contains very rich lexical material. The electronic version was obtained from the publishers in the T602 format (T602 is a Czech text editor). The book can be used solely for scientific and academic purpose.

As computed by the Unix program `wc` over the whole CES-1 document, the Czech Fiction corpus contains 94965 words in 1146353 Bytes.

3.2.2 Structure of the Corpus

The corpus consists of 7 <**div type=chapter**> elements. After 2 introductory chapters further divided into sections, the chapters 3 through 7 contain descriptions of the most famous European opera composers which is structurally expressed as <**div type=composer**>. Thus, chapter 3 contains 13 composers, chapter 4 3 composers, chapter 5 contains 15 composers, chapter 6 has 22 composers and chapter 7 contains the descriptions of the work of 4 composers. The division to chapters is made according to the nationality principle - the opera composers belonging to the same nation are included in the same chapter. Thus, the chapters are entitled Italská opera, Francouzská opera etc. Each composer unit is further subdivided into the short introduction about the composer in question and then follow the descriptions of his operas in <**div type=opera**>. Each composer (opera) is introduced by the **head** attribute which is then followed by the description of the composer (opera). Each opera is further divided into individual acts.

The **<div type=chapter>** elements have the **n** attribute, giving the chapter number, and the **id** attribute, whose value has the prefix **OPERA** followed by a period and the chapter number, e.g. **<div type=chapter n=3 id=OPERA.3>** Further in-depth structure is also reflected by successive numbers hierarchically following the book identification OPERA.

On the lowest structural level the text is composed of paragraphs containing subparagraph level tags, such as **<q>**, **<name>**, **<num>**, **<abbr>**, **<foreign>** etc.

Rendering information is given as the CES conformant two-letter value of the **rend** attribute. Each dash except within the values of other tags is represented by the SGML entity `mdash`.

Example from the corpus:

```
<div type=chapter n=3 id=OPERA>
<body lang=cs id=OPERA>
<div type=chapter n=1 id=OPERA.1>

<head>
<hi rend=ca>&Uacute;VOD</hi>
</head>

<p>
Obliba opern&iacute;ho &uacute;tvaru byla u n&aacute;s oded&aacute;vna
mimo&rcaron;&aacute;dn&aacute;, ale v
posledn&iacute; dob&ecaron; se na&scaron;e obecenstvo s operou
s&zcaron;&iacute;v&aacute; st&aacute;le v&iacute;ce. Stala
se nejsch&uring;dn&ecaron;j&scaron;&iacute; cestou k proniknut&iacute; do
sv&ecaron;ta hudebn&iacute; kr&aacute;sy. Jsou
v&scaron;ak &uacute;skal&iacute;, kter&aacute;
za&ccaron;&iacute;naj&iacute;c&iacute;mu z&aacute;jemci
p&uring;sob&iacute; nesn&aacute;ze. Pat&rcaron;&iacute; k
nim slo&zcaron;itost opern&iacute;ho &uacute;tvaru,
obt&iacute;&zcaron;n&aacute; zejm&eacute;na p&rcaron;i
sledov&aacute;n&iacute;
opery poprv&eacute; vid&ecaron;n&eacute; &ccaron;i dokonce, z rozhlasu jen
sly&scaron;en&eacute;. Oded&aacute;vna
proto vznikaly pom&uring;cky k snadn&ecaron;j&scaron;&iacute;mu
pochopen&iacute; pr&uring;b&ecaron;hu d&ecaron;je i
hudby opern&iacute;ho d&iacute;la.
</p>

<p>
Prvn&iacute; &ccaron;eskou informa&ccaron;n&iacute; knihou byla
<q rend="PRE lsquo POST rsquo">
<name>&Ccaron;esk&aacute; zp&ecaron;vohra brat&rcaron;&iacute;
Horn&uring;</name>
</q>,
vydan&aacute; roku
<date><num>1903</num></date>
u
<name type=org>Grossmanna a Svobody</name>
v
<name type=place>Praze</name>. V abecedn&iacute;m se&rcaron;azen&iacute;
skladatel&uring;
uv&aacute;d&ecaron;la stati o &ccaron;esk&yacute;ch oper&aacute;ch od
&Scaron;kroupova
<q rend="PRE lsquo POST rsquo">
<name>Dr&aacute;ten&iacute;ka</name>
```

```
</q>
a&zcaron; po d&iacute;la proveden&aacute; do roku
<date><num>1902</num></date>.
Ov&scaron;em jejich kn&iacute;&zcaron;ka brzy nesta&ccaron;ila. Roku
<date><num>1910</num></date>
napsal
<name type=person>A. Tvrdek</name> knihu, kterou pod n&aacute;zvem
<q rend="PRE lsquo POST rsquo">
<name>Anthologie z oper</name>
</q> vydal
<name type=person>Emil &Scaron;olce</name> v
<name type=place>Tel&ccaron;i</name>.
V n&iacute; jsou obsa&zcaron;ena u&zcaron; vedle z&aacute;kladn&iacute;ch
&ccaron;esk&yacute;ch oper i ta ciz&iacute;
d&iacute;la, kter&aacute; se tehdy hr&aacute;la na &ccaron;esk&yacute;ch
jevi&scaron;t&iacute;ch. Dal&scaron;&iacute; publikac&iacute;
tohoto typu pak byl
<q rend="PRE lsquo POST rsquo">
<name>Sv&ecaron;t v ope&rcaron;e</name>
</q>
(
<date><num>1934</num></date>)
sepsan&yacute;
<name type=person>J. Branbergem</name> a
<name type=person>Z. M&uuml;nzerovou</name>.
Podruh&eacute; titul vy&scaron;el roku
<date><num>1939</num></date>
a pot&rcaron;et&iacute;, roz&scaron;&iacute;&rcaron;en i o &rcaron;adu
balet&uring; na v&iacute;ce ne&zcaron;
<num>600</num> v&yacute;klad&uring; d&ecaron;j&uring;, roku
<date><num>1947</num></date>.
</p>
```

3.2.3 Structure of the Original

The original was technically organised according to printing columns (approx. 400 columns). They were omitted in our subsequent markup. Logical structure of the book was meticulously respected. Paragraph boundaries were marked in the original and rendition information in the original was respected in the markup.

3.2.4 Markup Process

The electronic version was at our disposal. It was semiautomatically marked up in a very detailed way (the corpus contains a lot of proper nouns which have all been marked by the <**name**> tag). Possessive adjectives derived from proper nouns have not been marked by the <**name**> attribute. Also some spelling errors were corrected during the markup. Finally, the whole corpus was validated by the nsgmls parser.

3.3 Estonian

Contributors: Heiki-Jaan Kaalep, Leho Paldre, Heili Orav, Urve Talvik and Kadri Muischnek

3.3.1 Description of the Corpus

The Estonian fiction corpus consists of 51 excerpts from Estonian novels or short stories from 1985. Each excerpt is approximately 2000 words long. The digital source used as the basis of encoding was provided by the University of Tartu, as an output of the project "Creating an Estonian text corpus". License agreements with the authors ensure that the texts are free to be distributed in any form for academic purposes.

The Estonian fiction corpus contains 104435 words, as indicated in the header of the tagged version.

3.3.2 Structure of the Corpus

The corpus body consists of 51 <**div type=excerpt**>. As there are excerpts from various places of the novels, all <**div**>-s below the level of "excerpts" are more or less arbitrary; e.g. the first <**div**> below some "excerpt" may well be <**div chapter=45**>.

The text is tagged up to the level of sentences. Names, abbreviations and direct speech have been tagged also.

Rendering information has been included with the appropriated tags.

Example from the corpus:

```
<div type="excerpt">
          <byline>
            <docauthor>Viivi Luik</docauthor>
            <title>Seitsmes rahukevad</title>
            <name type="org">Eesti Raamat</name>
            <name type="place">Tallinn</name>
            <date>1885</date>
            <num type="pages">5-8</num>
          </byline>
<p>
<s>
Suurtes ja hallides tee&auml;&auml;rsetes taludes olid elanud kulakud ja
rauds&auml;ngijalgadesse kulda peitnud.</s>
<s>
&Uuml;he talu perenaine oli aga ennast koguni s&auml;ngijala k&uuml;lge
&auml;ra poonud.</s>
<s>
M&otilde;ned lagunenud s&auml;ngid vedelesid veel praegugi n&otilde;gestes.</s>
</p>
```

3.3.3 Structure of the Original

The original came from the project "Creating an Estonian text corpus" in the form of electronic versions of 2000-word excerpts of Estonian novels and short stories, tagged to the sentence level in a TEI-like manner. The Estonian diacritics were encoded in extended ASCII. Every excerpt was a separate file.

An example of the original follows:

```
Viivi Luik "Seitsmes rahukevad" 1985, lk.3-8.
<p>
        <s>Suurtes ja hallides tee=84=84rsetes taludes olid elanud
kulakud ja rauds=84ngijalgadesse kulda peitnud.</s>
<s>=9Ahe talu perenaine oli aga ennast koguni s=84ngijala k=81lge
=84ra poonud.</s>
<s>M"ned lagunenud s=84ngid vedelesid veel praegugi n"gestes.</s>
</p>
```

3.3.4 Markup Process

The TEI-like texts were automatically converted to CES-tagged versions by a script written by a student of linguistics of the University of Tartu, Leho Paldre. The result was hand-validated. The separate files were then united into one file and modified to be CES-type as a single document.

3.4 Hungarian

Contributors: Csaba Oravecz and Júlia Pajzs (RIL)

3.4.1 Description of the Corpus

The contents of the Hungaian MULTEXT-East fiction corpus is four excerpts from four 20th century Hungarian novels. These are "Számdás" and "Falusi krónika" by Péter Veres, published in 1937 and 1944, respectively; "Budai oroszlán" by István Sőtér, published in 1978; and "Égető Eszter" by László Németh, published in 1956. There are approximately 50 pages from each novel in the corpus.

The digital source of the encoding was the National Corpus for the Hungarian Historical Dictionary. Since the distributor for this corpus is also the Research Institute for Linguistics, no licence agreement was necessary to allow its use. The copyright of the original printed versions is settled with three sources, for the fourth, verbal agreement is obtained, written licence is to be received.

The Hungarian Fiction Corpus now has 71777 words.

3.4.2 Structure of the Corpus

The corpus body consists of 4 <**div type=excerpt**>, each of which starts off with a <**head**> giving under <**byline**> the title, the author, the publisher and the publishing date of the novel from which the excerpt in question is taken. The number of pages appearing in the corpus is also indicated. <**div type=excerpt**> is sub-divided into <**div type=section**>, denoted in the printed editions by starting the paragraph on a new page.

The <**div type=section**> elements have the **n** attribute, giving the section number.

The text is segmented into paragraphs, with the <**quote**> element marked-up at the paragraph level.

Sub-paragraph tagging consists of <**abbr**>, <**foreign**>, <**hi**> and <**q**>. Direct speech has been marked-up by <**q**> even where there is no typographical marking in the printed text. Names with type attribute are only marked up in the sample version.

Rendering information is given in the same way as in the Hungarian CES1 version of “1984”.

Here follows an example from the corpus:

```
<mentioned rend="PRE ldquor POST rdquor">magyar</mentioned>-t
l&aacute;tta benn&uuml;nk (sz&eacute;p,
de ritka volt ez az igazi nemzeti &eacute;rz&eacute;s) &eacute;s az &uuml;gy&uuml;nket
a maguk&eacute;v&aacute; tett&eacute;k. De &eacute;ppen ez lett a
baj. Mert mi ki akartunk sz&aacute;llni j&oacute;val
<name type=place>Nagyv&aacute;rad</name> el&odblac;tt s &odblac;k biztattak, hogy a
v&aacute;rad-velencei kis&aacute;llom&aacute;son sincsen &odblac;rs&eacute;g.
</p>
<p>
Az asszonyka &eacute;s a kalauz az ajt&oacute;b&oacute;l lest&eacute;k, mi meg az
ablakon, hogy van-e vizsg&aacute;lat. S amikor behaladt a vonat &eacute;s
megl&aacute;ttuk a szuronyokat, megijedt&uuml;nk.
<q rend="PRE mdash POST mdash">Na most belement&uuml;nk
a farkasverembe </q> gondoltuk &eacute;s hamarj&aacute;ban nem is tudtuk, hogy
```

3.4.3 Structure of the Original

The electronic source was a not entirely SGML conformant encoded ASCII file, with Hungarian characters being encoded in a special ‘Prószéky’-code. Paragraph boundaries were marked, and additional sub-paragraph information was also included, such as foreign words and highlighted text marked. For a sample see below:

```
<l></par></l>
<l><par>Amikor azta1n ala1i1rta egy-egy csapat a szerzo3de1st, akkor</l>
<l>bementek a kocsma1ba. Minden csapat a megszokott kocsma1ja1ba.</l>
<l>Az u1jvila1giak a ,,kis zsido1''-hoz, a ba1nlakosiak Zsiga1hoz,</l>
<l>ke1so3bb Spitzerhez. </par></l>
<l><par>A bandagazda felva1ltotta a nagy pe1nzeket e1s elosztotta.</l>
<l>Egy-egy forint ,,felpe1nz'' ja1rt mindenkinek. Persze, mindenki</l>
<l>rendelt legala1bb egy fe1ldecit, ez ma1r a tisztesse1g </l>
</page>
```

3.4.4 Markup Process

The electronic source was first automatically converted into a quasi-sgml encoding. This version was then additionally corrected and marked up to CES1 conformance by hand. Name tagging in the sample was carried out automatically, with manual correction. Since in Hungarian printing the typographical rendering of direct speech overlaps with the typographical rendering of other types of text, marking with <**q**> tags had to be done manually. The corpus text was also cross-checked against the printed editions.

3.5 Romanian

Contributors: Dan Tufiş and Ştefan Bruda (RACAI), Lidia Diaconu, Călin Diaconu (ICI)

3.5.1 Description of the Corpus

The contents of the Romanian MULTEXT-East fiction corpus is made of three novels "Flăcări sub cruce", "Obreja" and "Testament între înger şi diavol" by Mihai Rădulescu. The novels were published by "Ramida" Publishing House (the first two novels) and "PIKA" Publishing House (the last one). The sources also included introductions by author or by H.H. Teodosie Snagoveanu.

The digital source used as the basis of encoding was provided by the "Ramida" Publishing House and "PIKA" Publishing House based on the author's written permission.

The Romanian site obtained a license agreement allowing the use of the three novels for the purposes of the MULTEXT-East project, signed by the author, Mihai Rădulescu.

As computed by the Unix program `wc` over the Romanian Fiction corpus has 160405 words (39389, 39521 and 81495 for each novel).

3.5.2 Structure of the Corpus

The corpus body consists of some <**div type=part**>, each of which starts off with a <**head**> giving the chapter title. Each part may be divided by some <**div type=chapter**>, which also begin with a <**head**>.

The <**div type=chapter**> and <**div type=part**> elements have the **n** attribute, giving the chapter or part number, and the **id** attribute, whose value has the prefixes conformant with the novel, e.g. <**div id="obreja.2.1" type="chapter"**>.

The text is segmented into paragraphs, with the <**head**>, <**quote**> and <**poem**> elements marked-up at the paragraph level.

Sub-paragraph tagging consists of <**hi**> and <**q**>. Direct speech has been marked-up by <**q**> even where there is no typographical marking to that effect in the printed text.

Rendering information, given as the CES conformant two-letter value of the **rend** attribute has been included with the appropriated tags and, for mdash and capitalisation, retained in the tag content.

The tag usage for the three novels is the following:

- “Flăcări sub cruce”

```
<tagsdecl>
  <tagusage gi=body occurs=1></tagusage>
  <tagusage gi=div occurs=18></tagusage>
  <tagusage gi=head occurs=20></tagusage>
  <tagusage gi=hi occurs=110></tagusage>
  <tagusage gi=l occurs=47></tagusage>
  <tagusage gi=p occurs=817></tagusage>
  <tagusage gi=poem occurs=5></tagusage>
  <tagusage gi=q occurs=564></tagusage>
  <tagusage gi=text occurs=1></tagusage>
</tagsdecl>
```

- “Obreja”

```
<tagsdecl>
  <tagusage gi=body occurs=1></tagusage>
  <tagusage gi=div occurs=17></tagusage>
  <tagusage gi=head occurs=32></tagusage>
  <tagusage gi=hi occurs=368></tagusage>
  <tagusage gi=l occurs=4></tagusage>
  <tagusage gi=p occurs=563></tagusage>
  <tagusage gi=poem occurs=1></tagusage>
  <tagusage gi=q occurs=821></tagusage>
  <tagusage gi=text occurs=1></tagusage>
  <tagusage gi=foreign occurs=1></tagusage>
</tagsdecl>
```

- “Testament între înger şi diavol”

```
<tagsdecl>
  <tagusage gi=body occurs=1></tagusage>
  <tagusage gi=div occurs=45></tagusage>
  <tagusage gi=head occurs=84></tagusage>
  <tagusage gi=hi occurs=338></tagusage>
  <tagusage gi=l occurs=171></tagusage>
  <tagusage gi=p occurs=1464></tagusage>
  <tagusage gi=poem occurs=12></tagusage>
  <tagusage gi=q occurs=808></tagusage>
  <tagusage gi=quote occurs=341></tagusage>
  <tagusage gi=text occurs=1></tagusage>
</tagsdecl>
```

Example from the corpus:

```
<div id="obreja.1" type="part">
 <head> AUZI-M&Abreve;, DOAMNE! </head>
 <head> -Cuv&acirc;nt c&abreve;tre cititor- </head>
```

```
<P>
Una dintre &icirc;ntreb&abreve;rile sf&acirc;&scedil;ietoare ce apar adesea
pe buzele
neferici&tcedil;ilor care trec prin &icirc;ncerc&abreve;ri mult prea grele
pentru puterile
lor este: <q rend=dblq> De ce m-a p&abreve;r&abreve;sit Dumnezeu? </q>
&Icirc;nso&tcedil;it&abreve; de strig&abreve;te nedumerite &scedil;i disperate,
speran&tcedil;a &icirc;n interven&tcedil;ia divin&abreve;
r&abreve;m&acirc;ne cu
at&acirc;t mai mare: <q rend=dblq> Unde e&scedil;ti, Doamne?! </q> Sau,
relu&acirc;nd
cuvintele psalmistului: <q rend=dblq> Auzi-m&abreve;, Doamne! </q> </P>
<P>
Omul are nevoie permanent&abreve; de P&abreve;rintele s&abreve;u. &Icirc;i
este greu
s&abreve; &icirc;n&tcedil;eleag&abreve; &icirc;nt&acirc;rzierea ajutorului
dumnezeiesc
l&abreve;murit, ori c&abreve; el se manifest&abreve; nev&abreve;zut &scedil;i
nerecunoscut &icirc;ntru &icirc;nt&abreve;rirea puterilor d&abreve;ruite celui
&icirc;ndurerat , pentru ca acesta s&abreve;-&scedil;i poat&abreve;
r&abreve;bda chinul. </P>
<P>
Una dintre misiunile Teologiei este s&abreve; fac&abreve; de
&icirc;n&tcedil;eles
Divinitatea. Alta - mult mai apropiat&abreve; de orizonturile &icirc;nguste ale
f&abreve;pturilor umane ce suntem &scedil;i de aceea mai important&abreve;
pentru
biata noastr&abreve; neputin&tcedil;&abreve; -, alt&abreve; misiune a
Teologiei este
s&abreve;-i explice suferitorului de ce este l&abreve;sat de Dumnezeu s&abreve;
se chinuiasc&abreve;, s&abreve;-i explice lui Iov de ce i s-a
&icirc;ng&abreve;duit
satanei s&abreve;-l pun&abreve; la &icirc;ncercare. </P>
<P>
.
.
.
```

3.5.3 Structure of the Original

The original versions of the three novels, which were the basis for the encoding, were ASCII exported files from an uncommon text-editor (chiwriter) with a relatively transparent encoding: paragraph boundaries were marked, and a few useful formatting codes were included. Due to the lack of printed versions, no hilighting marking has been provided, except for the "rend=dblq" marking found in the text. Also, the text contained a number of typo errors, which were also in the printed version. We corrected these errors in the fiction corpus.

3.5.4 Markup Process

Due to the lack of printed versions, no hilighting marking has been provided, except for the "rend=dblq" marking found in the text. Also, the text contained a number of typo errors, which were also in the printed version. We corrected these errors in the fiction corpus.

3.6 Slovene

Contributors: Tomaž Erjavec (IJS) and Miro Romih (Amebis d.o.o.)

3.6.1 Description of the Corpus

The contents of the Slovene MULTEXT-East fiction corpus is the novel "Galjot" by Drago Jančar, which tells of the wanderings of J. Ot from the German duchy of Neisse in the Slovene counties and the Mediterranean in the 17th century. The novel was first published by "Pomurska založba", Murska Sobota in 1978, with the last publication being the third, corrected edition by "Mladinska knjiga", Ljubljana, in 1984. The digital source used as the basis of encoding was provided, via Amebis d.o.o., by the Slovene Society for Blind and Visually Impaired, where the 1984 printed edition was OCRed. This source also included a short introduction by M. Kramberger, which was not retained in the Slovene Fiction MULTEXT-East corpus.

The Slovene site obtained a license agreement allowing the use of "Galjot" for the purposes of the MULTEXT-East project, signed by the author, Drago Jančar.

As computed by the Unix program `wc` over the whole CES-1 document, the Slovene Fiction corpus has 101308 words.

3.6.2 Structure of the Corpus

The corpus body consists of 25 <**div type=chapter**>, each of which starts off with a <**head**> giving the chapter number and an <**opener**> with a short expressionistic summary of the chapter. Chapters are further sub-divided into <**div type=section**>, denoted in the printed edition by increased spacing only — some of these divisions also begin with a <**head**>.

The <**div type=chapter**> elements have the **n** attribute, giving the chapter number, and the **id** attribute, whose value has the prefix **galjot** followed by a period and the chapter number, e.g. <**div type=chapter n=2 id=galjot.2**>

The text is segmented into paragraphs, with the <**head**>, <**quote**>, <**opener**>, <**list**>, <**poem**> and <**table**> elements marked-up at the paragraph level.

Sub-paragraph tagging consists of <**abbr**>, <**foreign**>, <**hi**> and <**q**>. Direct speech has been marked-up by <**q**> even where there is no typographical marking to that effect in the printed text.

Rendering information, given as the CES conformant two-letter value of the **rend** attribute has been included with the appropriated tags and, for mdash and capitalisation, retained in the tag content.

Example from the corpus:

```
<div type=chapter n=1 id=galjot.1>
<head>
1
</head>

<opener rend="IT">
```

```
Gosti skladi zraka. Sluz se vzpenja po stenah. Prihod iz mo&ccaron;virja.
Ku&zcaron;ni komisarji v de&zcaron;eli. Tak &ccaron;uden vstop, tak vinski
za&ccaron;etek.
</opener>

<div type=section>
<p>
Temne lise vlage so se spakovale po zidu. Zdelo se mu je, da v tej gluhi
tihoti lezejo skupaj in narazen in da s svojim neznansko po&ccaron;asnim
gibanjem oblikujejo nedolo&ccaron;ljive podobe. Spodaj je bilo okrog in okrog
mokro, zid je bil prav do &ccaron;rnega prepojen s sluzasto vodno
```

3.6.3 Structure of the Original

The OCR'ed version, which was the basis for the encoding, was an ASCII file with a relatively transparent encoding: paragraph boundaries were marked, and a few useful formatting codes were included. However, the text contained a number of recognition errors, e.g. typos and misrendered punctuation. A sample of the original is given below.

```
#s   #+1#-
#s2   #+Gosti skladi zraka. Sluz se vzpenja po stenah. Prihod iz mo~virja. Ku`ni
komisarji v de`eli. Tak ~uden vstop, tak vinski za~etek.#-
#s2   #+Temne lise vlage so se spakovale po zidu. Zdelo se mu je, da v tej gluhi
tihoti lezejo skupaj in narazen in da s svojim neznansko po~asnim gibanjem
oblikujejo nedolo~ljive podobe. Spodaj je bilo okrog in okrog mokro, zid je bil
prav do ~rnega prepojen s sluzasto vodno snovjo. Nevidno gibanje se je pomikalo
```

3.6.4 Markup Process

The OCR text was first automatically converted, by Amebis d.o.o., into a quasi-sgml encoding, and at the same time corrected for typos with a spelling-checker. This version has been then additionally corrected and marked up to CES1 conformance by IJS. As there are often no typographical marks which indicate direct speech, the <**q**> tags were inserted manually. The corpus text was also cross-checked against the printed edition of the novel.

Chapter 4

Multilingual Comparable 2: Newspapers

This chapter gives the details of the 'newspapers' part of the MULTEXT-East multilingual comparable corpus. As the name suggests, this set consists of articles from daily newspapers in the languages of the project. The composition of this set is as follows:

MULTEXT-East Multilingual Comparable Corpus, Newspapers:

- Bulgarian: "Kontinent" daily
- Czech: "Lidové noviny" daily
- Estonian: articles from 11 newspapers
- Hungarian: "Magyar Hirlap" daily
- Romanian: "Romånia Liberà" daily
- Slovene: "Dnevnik" daily

The following sections give the details for each of the above language components in turn.

4.1 Bulgarian

Contributors: Institute of Mathematics, Bulgarian Academy of Sciences

4.1.1 Description of the Corpus

The corpus contains about 150 articles (97256 words) from the daily newspapers "Capital", "Continent", "Pari" and "Standart".

The corpus was mainly obtained from the publishers of "Pari" and "Standart". This was done on a "friendship" basis — there was no official agreement between the Institute of Mathematics and the Publishing houses. Due to this, we had not much choice on selecting appropriate articles.

Moreover, part of the articles do not contain the name of their author. In such cases we have marked the publisher as an author. The dates when the articles were published were not known for some of the articles, so no information is given in these cases. Actually we did not have half of the electronic texts as published paper.

A small part of the corpus was typed-in (excerpts from the newspapers "Continent" and "Capital"). These texts are marked to the level of sentence (sentence excluded).

4.1.2 Structure of the Corpus

The corpus consists of 6 supersections, <**div type=supersection**> corresponding to the different sources of articles. Each supersection is identified with **id=Newsbg.N**, **N**=1–6.

Each supersection consists further of articles, some of which are grouped into sections (pertaining to a distinct topic, e.g. Politics). Each <**div**> has an **id** attribute with subsequent numbering.

The <**div type=article**> and optionally <**div type=section**> usually starts off with one or more <**head**> tags, giving the headline(s).

Pictures accompanying the articles are tagged by <**figure**> The attached text and the author are tagged by <**caption type=attached**> and <**caption type=byline**>. <**gap**> is used to indicate that the figure itself is not there. The place of a figure in the text is given using <**ptr**>.

If a picture is given in the newspaper alone - not accompanying an article, it is given in <**div type=frame**>. It may have <**head**> and <**caption**> for the text, attached to the picture.

Small pieces of information are tagged with <**div type=brief-news**>, <**head**> is optional for them. Usually such pieces in the newspapers are presented in a group. This is accounted for by the <**div type=chronicles**>.

The following is an example from the Bulgarian Newspaper corpus; the characters are, in the original, encoded as SGML ISO Cyrillic entities (e.g. `о`); to make the following sample more readable, they were automatically transliterated into Latin (e.g. `о` ⇒ `o`).

```
<div type=frame id=Newsbg.2.1.2>
<figure>
<gap>
<caption type=attached>
Siediem pozharni koli i diesietki ogniebortsi gasikha v
prodhardlzhieniie na chas vchiera pozhar na pokriva na koopieratsiya
na
<address>
<abbr>ul.</abbr> <hi rend=dblq>San
Stiefano</hi> <num>17</num>
</address> v stolitsata.
<q>Dnies bieshie prokliet dien, gasikhmie na tri miesta v
<name type=place>Sofiya</name></q>,
kazakha pozharnikari. Do khardsniya sliedobied pozharitie v stranata byakha
<num>12</num>, informirakha
ot
<name type=org>Natsionalnata
sluzhba za protivopozharna
okhrana</name>.
```

```
</caption>
<caption type=byline>Snimka
<docAuthor>Vasil
MILANOV</docAuthor>
</caption>
</figure>
</div>
```

4.1.3 Structure of the Original

The original electronic files were made by the newspaper editors. The provided texts contained usually preliminary versions of the articles and in general no identification information (date, author) This information was supplied by us comparing the text and the published newspaper.

The following is an example from the original digital source; the characters are, in the original, encoded in an 8-bit Cyrillic font — to make the following sample readable, they were automatically transliterated into Latin.

```
Siediem pozharni koli i diesietki ogniebortsi gasikha v
prodhardlzhieniie na chas vchiera pozhar na pokriva na koopieratsiya
na ul. San Stiefano 17 v stolitsata.  Dnies bieshie prokliet dien,
gasikhmie na tri miesta v Sofiya, kazakha pozharnikari.  Do khardsniya
sliedobied pozharitie v stranata byakha 12, informirakha ot
Natsionalnata sluzhba za protivopozharna okhrana.

Snimka Vasil
MILANOV
```

4.1.4 Markup Process

The texts have been first converted to ASCII texts, then spell-checked and all “strange” symbols removed. The texts have been marked up to CES1 conformance manually. As the nsmgl editor could not be started, programs have been written to count the tags, to do syntactic validation, clean the text of unnecessary spaces, etc.

However in its most, the markup process was manual.

4.2 Czech

Contributors: Vladimír Petkevič, Věra Schmiedtová (UJČ AVČR) and Jana Klímová (FFUK)

4.2.1 Description of the corpus

The contents of the Czech MULTEXT-East newspaper corpus is formed by 451 articles from the Czech daily newspaper Lidové noviny from the period 1991-1994 (some articles published in this period were chosen). The newspaper texts were available on the basis of the written contract.

The newspaper data coming into MULTEXT-East were already automatically marked up. This markup used for the texts in the Czech National Corpus which is different from the one required by CES was automatically transformed to something near CES. Then semiautomatic and hand validation had to be performed so as the markup were CES conformant. The main bulk of the corpus is formed by short articles from various areas ranging from politics over interviews to sport. When choosing the articles for the MULTEXT-East corpus, no specific guidelines were respected. The articles reflect a usual newspaper style. On the basis of the contract with the publisher Lidové noviny the newspaper data can be used solely for scientific and academic purposes.

As computed by the Unix program `wc` over the whole CES-1 document, the Czech Newspaper corpus has 99564 words in 1210258 bytes.

4.2.2 Structure of the corpus

The corpus body consists of 451 <**div type=article**>, each of which contains one article from the original newspaper data. The topmost segmentation is very flat. Most articles are introduced by one or more <**head**>(s) specifying the headline(s) or by one <**opener**>. The article is composed of paragraphs. The whole article is mostly ended by <**byline**> giving the name of the document author. Document authors are included where they appeared in the original, usually at the end of the articles; they are marked-up as:

<**byline**><**docauthor**>Author</**docauthor**></**byline**>.

The <**div**> elements have the **n** attribute, giving the successive number of the article and the **id** attribute. Articles are not further divided into article parts by a <**div**> tag.
Example from the corpus:

```
<div complete=y type=article n=1 id=NEWS.1>

<head>
 Zpracovatel&uring;m druhotn&yacute;ch surovin hroz&iacute; krach
</head>

<p>
<hi rend=bo>

&Uacute;pln&yacute;m zastaven&iacute;m dovozu druhotn&yacute;ch
surovin ohrozilo ministerstvo &zcaron;ivotn&iacute;ho
prost&rcaron;ed&iacute; podnikatele, kte&rcaron;&iacute; se
zam&ecaron;&rcaron;ili na ekologick&eacute; technologie.
Proto&zcaron;e se u n&aacute;s tak&rcaron;ka nepoda&rcaron;ilo
zav&eacute;st t&rcaron;&iacute;d&ecaron;n&iacute; odpad&uring;,
&rcaron;ada v&yacute;robc&uring; je na dovozu druhotn&yacute;ch
surovin z&aacute;visl&aacute;. Na p&rcaron;&iacute;m&eacute;
nebezpe&ccaron;&iacute; &uacute;padku
<abbr>LN</abbr>
 upozornila &ccaron;eskolipsk&aacute; firma
<name type=org>Presta</name>
, zpracov&aacute;vaj&iacute;c&iacute; polystyren.
</hi>
</p>
```

```
<p>

Hromadn&yacute; z&aacute;kaz dovozu byl motivov&aacute;n snahou
ministerstva, aby se
<abbr>&Ccaron;R</abbr>
 nestala
<q>smeti&scaron;t&ecaron;m
<name type=place>Evropy</name>
</q>.

Minister&scaron;t&iacute; &uacute;&rcaron;edn&iacute;ci toti&zcaron;
dosud povolovali i dod&aacute;vky odpad&uring;, z nich&zcaron; byla
vyu&zcaron;iteln&aacute; jen &ccaron;&aacute;st a zbytek se hromadil
na na&scaron;ich skl&aacute;dk&aacute;ch nebo p&rcaron;isp&ecaron;l
zne&ccaron;i&scaron;t&ecaron;n&iacute; ovzdu&scaron;&iacute;
p&rcaron;i spalov&aacute;n&iacute;. Pod pl&aacute;&scaron;t&iacute;kem
dovozu &zcaron;elezn&eacute;ho &scaron;rotu bylo
<abbr>nap&rcaron;.</abbr>
povoleno dov&aacute;&zcaron;et &scaron;pony,
zne&ccaron;i&scaron;t&ecaron;n&eacute;
<abbr>PCB</abbr>.
 Odpov&ecaron;dn&yacute; pracovn&iacute;k ministerstva
<abbr>ing.</abbr>
<name type=person>Durd&iacute;k</name>
 v&scaron;ak nebyl schopen vysv&ecaron;tlit, pro&ccaron; nen&iacute;
povolen dovoz pou&zcaron;it&eacute;ho polystyrenu ze
<abbr>SRN</abbr>,

kdy&zcaron; se u n&aacute;s zpracov&aacute;v&aacute; bez
jedin&eacute;ho gramu odpadu a je <q>&ccaron;ist&scaron;&iacute;</q>
ne&zcaron; mnoh&aacute; p&rcaron;&iacute;rodn&iacute; surovina.
P&rcaron;esto&zcaron;e se po st&iacute;&zcaron;nostech
<name>Presty</name>
do &rcaron;e&scaron;en&iacute; probl&eacute;mu anga&zcaron;oval i
premi&eacute;r
<name type=person>Klaus</name>
, povolen&iacute; ve sl&iacute;ben&eacute;m term&iacute;nu
<abbr>M&Zcaron;P</abbr>
 nevydalo.
</p>
```

4.2.3 Structure of the original

The original text for the MULTEXT-East corpus was already marked up according to the DTD used for the Czech National Corpus. Simple transformation was used to get near the CES needed for the ME files.

Original rendition is mostly reflected in the CES markup. Single quotes are used for direct speech in the newspapers.

4.2.4 The markup

The subparagraph markup is relatively detailed, the <**date**>, <**num**>, <**abbr**> and <**q**> are primarily used. However, no name tags are used. During markup many typos were corrected.

The whole corpus was finally validated by nsgmls validator.

4.3 Estonian

Contributors: Heiki-Jaan Kaalep, Heili Orav and Urve Talvik

4.3.1 Description of the Corpus

The contents of the Estonian MULTEXT-East newspaper corpus are over 300 articles from eleven newspapers from 1985. The digital source used as the basis of encoding was provided by the University of Tartu, as an output of the project "Creating an Estonian text corpus". Licence agreements with the original publishers of the newspapers ensure that the texts are free to be distributed in any form for academic purposes.

The Estonian version of the newspaper corpus contains 112002 words, as indicated in the header of the tagged version.

4.3.2 Structure of the Corpus

The corpus body consists of 11 <**div type=newspaper**>, each of which contains several files from the original digital data, constituting several articles. Altogether there are over 300 <**div type=article**>. In fact the corpus contains besides "real" articles also announcments and news from news agencies.

The text is segmented up to the sentence level. Names, lists, numbers, abbreviations and direct speech are also tagged.

Rendering information attribute has been in most cases included with the appropriated tags.

Example from the corpus:

```
        <div type="newspaper">
          <byline>
            <docauthor></docauthor>
            <title>&Otilde;htuleht</title>
            <name type="org">EKP KK kirjastus</name>
            <name type="place">Tallinn</name>
            <date>25/04/1985</date>
            <num type="issue">95</num>
            <num type="pages">1-4</num>
          </byline>
       <div type="article">
          <byline>
            <docauthor></docauthor>
            <title>V&otilde;tame v&auml;&auml;riliselt vastu NLKP XXVII
                   kongressi!</title>
            <num type="page">1</num>
          </byline>
      <head>
V&otilde;tame v&auml;&auml;riliselt vastu<name>
```

```
<abbr expan="N&otilde;ukogude Liidu Kommunistlik Partei">
NLKP</abbr>

XXVII
kongressi</name>
!</head>
<p>
<s>
<hi rend=bold>

Linna asutustes ja ettev&otilde;tetes tutvutakse<name>
<abbr expan="N&otilde;ukogude Liidu Kommunistlik Partei">
NLKP</abbr>
Keskkomitee aprillipleenumi</name>
materjalidega ning<name type=org>
Keskkomitee</name>
peasekret&auml;ri<name type=person>
Mihhail Gorbat&scaron;ovi</name>
<name>
ettekandega
<abbr expan="N&otilde;ukogude Liidu Kommunistlik Partei">
NLKP</abbr>
korralise,
XXVII
kongressi
kokkukutsumisest ning selle ettevalmistamise ja l&auml;biviimisega seotud
&uuml;lesannetest
</name>
.</hi>
</s></p>
```

4.3.3 Structure of the Original

The digital source used as the basis of encoding was provided by the University of Tartu. It consisted of data files, one file per article and tagged according to TEI. Each file starts off by header information, followed by the text.

```
<!DOCTYPE TEI.2 SYSTEM 'tei2.dtd'[
<!ENTITY % TEI.general 'INCLUDE'>
<!ENTITY % TEI.analysis 'INCLUDE'>
<!ENTITY % TEI.figures 'INCLUDE'>
<!ENTITY % ISOLat1 PUBLIC "ISO  8879-1986//ENTITIES Added Latin 1//EN">
%ISOLat1;
<!ENTITY % ISOLat2 PUBLIC "ISO 8879-1986//ENTITIES Added Latin 2//EN">
%ISOLat2;
<!ENTITY % ISOnum  PUBLIC "ISO 8879-1986//ENTITIES Numeric and Special Graphic//EN">
%ISOnum;
<!ENTITY % ISOpub  PUBLIC "ISO 8879-1986//ENTITIES Publishing//EN">
%ISOpub;
<!ENTITY % ISOdia  PUBLIC "ISO 8879-1986//ENTITIES Diacritical Marks//EN">
%ISOdia;
<!ENTITY % MYent   PUBLIC "ISO 8879-1986//ENTITIES Characters Unfindable//EN">
%MYent;
```

```
]>
<tei.2 lang=ET>
<teiheader>
   <filedesc>
      <titleStmt>
         <title>TEI version of: V&otilde;tame v&auml;&auml;riliselt vastu NLKP
                XXVII kongressi!</title>
         <author></author>
         <principal>Heiki-Jaan Kaalep</principal>
         <respStmt>
            <resp>entered the text</resp>
            <name>Riina Mosna</name>
            <resp>validated with psgml</resp>
            <name>Heili Orav</name>
            <resp>validated with sgmls</resp>
            <name>Leho Paldre</name>
            <resp>finalised the header</resp>
            <name>Heiki-Jaan Kaalep</name>
         </respStmt>
      </titleStmt>
      <extent>8396 bytes</extent>
      <publicationStmt>
         <authority>T&Uuml; arvutuslingvistika uurimisgrupp</authority>
         <pubPlace>Tartu, Tiigi 78-232</pubPlace>
         <date>Fall 1995</date>
         <availability>
            <p>Available with prior consent for purposes of research only
         </availability>
      </publicationStmt>
      <sourcedesc>
         <bibl>
            <title level=a>V&otilde;tame v&auml;&auml;riliselt vastu NLKP XXVII
                   kongressi!</title>
            <title level=m>&Otilde;htuleht</title>
            <title level=s></title>
            <author></author>
            <biblScope type=issue>95</biblScope>
            <biblScope type=pages>pp. 1</biblScope>
            <imprint>
               <publisher>EKP KK Kirjastus</publisher>
               <pubPlace>Tallinn</pubPlace>
               <date>85-0-0</date>
            </imprint>
        </bibl>
      </sourcedesc>
   </filedesc>
   <encodingdesc>
        <projectdesc>
                <p>Estonian written text corpus of 1 million words
                   based on published texts from 1983-87
        </projectdesc>
        <samplingdecl>
                <p>Every text contains 1 whole unit of paper-printed text
                   (approximately of the size of 2,000 words unless the unit was smaller)
        </samplingdecl>
        <editorialdecl>
                <correction><p>No corrections</correction>
```

```
                <hyphenation><p>No hyphenated words in this electronic version.
                </hyphenation>
                <normalization><p>No normalization</normalization>
                <interpretation>
                   <p>Abbreviations are marked and their expanded form given.
                   <p>Proper named are marked with their type.
                   <p>All highlighted text is marked without interpretaton.
                   <p>Numbers are marked with numerical interpretation and type.
                   <p>Dates and time are marked with numerical interpretation.
                   <p>Lists are interpreted as clauses and items in list as phrases.
                <quotation marks=all><p></quotation>
                <segmentation>
                   <p>Up to the level of sentences</segmentation>
      </editorialdecl>
      <tagsdecl>
       <tagUsage gi=abbr occurs=4></tagUsage>
         <tagUsage gi=body occurs=1></tagUsage>
         <tagUsage gi=div0 occurs=1></tagUsage>
         <tagUsage gi=head occurs=1></tagUsage>
         <tagUsage gi=hi occurs=7></tagUsage>
         <tagUsage gi=name occurs=18></tagUsage>
         <tagUsage gi=num occurs=9></tagUsage>
         <tagUsage gi=p occurs=7></tagUsage>
         <tagUsage gi=q occurs=4></tagUsage>
         <tagUsage gi=s occurs=30></tagUsage>
         <tagUsage gi=text occurs=1></tagUsage>
      </tagsdecl>
   </encodingdesc>
   <profiledesc>
      <langUsage>
         <language id=ET>Estonian</language>
         <language id=DE>German</language>
         <language id=LA>Latin</language>
         <language id=EN>English</language>
         <language id=FR>French</language>
   </profiledesc>
</teiheader>
<text>
<body>
<div0 type=unknown><head>
V&otilde;tame v&auml;&auml;riliselt vastu<name  type=event>
<abbr expan="N&otilde;ukogude Liidu Kommunistlik Partei">
NLKP</abbr>
<num type=Roman value=" 27 ">
XXVII</num>
kongressi</name>
!</head>
<p>
<hi rend=bold>
<s>
Linna asutustes ja ettev&otilde;tetes  tutvutakse<name type=event>
<abbr expan="N&otilde;ukogude Liidu Kommunistlik   Partei">
NLKP</abbr>
Keskkomitee aprillipleenumi</name>
materjalidega ning<name type=org>
Keskkomitee</name>
peasekret&auml;ri<name type=person>
```

```
Mihhail Gorbat&scaron;ovi</name>
<name type=product>
ettekandega<hi rend="PRE laquo POST raquo">
<abbr expan="N&otilde;ukogude Liidu Kommunistlik   Partei">
NLKP</abbr>
korralise,<num type=Roman  value=" 27 ">
XXVII</num>
kongressi
kokkukutsumisest  ning selle ettevalmistamise ja l&auml;biviimisega seotud
&uuml;lesannetest</hi>
</name>
.</s>
</hi></p>
<p><hi>
```

4.3.4 Markup Process

A script was written by Leho Paldre, a student of Linguistics at the University of Tartu to convert from TEI to CES1. As the DTD for CES changed some times during the first year of the project, the script had to be rewritten as many times. Some cases which the script either did not cover or processed incorrectly were tagged by hand.

Part of the sub-paragraph mark-up was deleted from the MTE version of the newspaper corpus because it is really not needed in MTE and validating it in the context of a changing DTD seemed to be a waste of time and resources.

4.4 Hungarian

Contributors: Csaba Oravecz, Tamás Váradi and Gábor (Kiss, RIL)

4.4.1 Description of the Corpus

The Hungarian newspaper corpus contains 205 articles from the daily newspaper "Magyar Hírlap". The articles are from the Jan 25 and Jan 31, 1996, issues. The digital source was provided by the Magyar Hírlap Publishing House Ltd. and was composed of data files with, if any, idiosyncratic encoding and embedded comments for the typesetters. The corpus represents a wide variety of articles each being characteristic of everyday journalism (i.e. long, essay type articles typically from supplements to weekend editions were not included).

A licence agreement was obtained allowing the use of these articles for the purposes of the MULTEXT-East project.

The Hungarian Newspaper Corpus now has 63351 words. The corpus is currently still under extension.

4.4.2 Structure of the Corpus

The corpus body consists of 2 <**div type=newspaper**>, each of which contains 94 and 111 <**div type=articles**>, respectively. In addition, a number of articles are grouped under one <**div**

type=storylist>, where they appeared as short snippets each having its own head in a coloumn under a common heading.

The <**div type=article**> usually begins with one or more <**head**> tags, giving the headline(s), and one <**byline**> representing the source the content of the article came from (typically a news agency). After this <**byline**> another head can potentially follow giving an abstract of the article.

Captions to the pictures accompanying the articles, when they were represented in the digital source, were also included in the corpus. They are normally given at the beginning of <**div type=article**>.

The <**div**> elements apart from the type attribute, have no other attributes.

Document authors are included where they appeared in the original, usually at the end of the articles; they are marked-up as <**byline**> <**docAuthor**>Author or Initials</**docAuthor**> </**byline**>.

The text is segmented into paragraphs, other paragraph level tagging is <**note**>. This is used when the article is continued on or from another page. In this case reference to the page is given between <**ref**>. However, no pointer is included.

Sub-paragraph tagging consists of <**abbr**>, <**name**>, and <**q**>. The former two were only marked up in the Jan 25 issue; semi-automatically, then were manually corrected and provided with the type attribute. <**q**> is given for the whole corpus.

The **rend** attributes on sub-paragraph tags are included in the same way as in the Hungarian version of "1984". No quotation marks are retained in the corpus.

Here follows an example from the corpus:

```
<byline><abbr>MH</abbr>-inform&aacute;ci&oacute;</byline>

<head>Az er&odblac;s havaz&aacute;s megb&eacute;n&iacute;totta a
<name type=place>Tasz&aacute;rr&oacute;l</name>, illetve
<name type=place>Kaposv&aacute;rr&oacute;l</name> indulni
k&eacute;sz&uuml;l&odblac; <name type=org>IFOR-konvojok</name>
mozg&aacute;s&aacute;t. Tegnap egyetlen g&eacute;pkocsiegys&eacute;g
sem tudott boszniai rendeltet&eacute;si helye fel&eacute; indulni.
</head>

<p>
A f&odblac;v&aacute;rosi <name type=org>Rend&odblac;ri
Ezred</name> rendfenntart&oacute;i &eacute;s a <name type=org>Somogy
Megyei Rend&odblac;r-f&odblac;kapit&aacute;nys&aacute;g </name>
```

4.4.3 Structure of the Original

The digital source used as the basis of encoding was provided by Magyar Hírlap Publishing House Ltd. It consisted of data files, one file per issue. Pratically all information regarding the actual layout of the text was encoded idiosyncratically by dint of a number of methods (special characters, line spacing, etc.). Some difference between the electronic text and the printed paper was at places observed. When possible, the printed version was considered as basis for the corpus encoding.

An example from the original (8 bit characters are not rendered):

```
MH-informci
Az ers havazs megbntotta a
Taszrrl, illetve Kaposvrrl indulni ksz*l
IFOR-konvojok mozgst. Tegnap egyetlen gpkocsiegysg sem
tudott boszniai rendeltetsi helye fel indulni.

A
fvrosi Rendri Ezred rendfenntarti s a Somogy
Megyei Rendr-fkapitnysg kzlekedsi rendrei
egsz nap vrtk, hogy a kt bzisrl felvezeti
felkrst kapjanak, de az - lapzrtnkig - elmaradt.
A
futak az *tinform dlutni tjkoztatja szerint
egyelre mindenhol jrhatak. A havazs a Dunntlon
a legintenzvebb, elssorban Zala s Vas megyben okoz
kzlekedsi gondokat. E ter*letek alsbbrend* tjain
flszlessgben szmtani kell hfvsokra is,
az orszg tbbi tjra ltalban a hkss,
latyakos felszn jellemz. Az autplykat
szzk, de a h nem vagy csak nagyon lassan olvad abban a
svban, amelyikben most ritkbban kzlekednek a
jrm*vek. Fennakadsokra a hegyes, dombos vidkeken lehet
szmtani. Megynknt mintegy 25 gp tiszttja az
utakat * tudatta az *tinform.
Budapesten is lelassult a
kzlekeds, az utak szsa nem hoz eredmnyt, mivel a
mnusz 5-7 fokos hidegben a hksa rfagy az
aszfaltra.
$$$

***
```

4.4.4 Markup Process

The original digital version was converted into an ASCII text version where possible extracting all information present, and converting it into CES1 conformant markup. However, given the rather unreliable methods of rendering layout information in the original, laborious correction process, done manually, was necessary to ensure conformance to the printed issues. During this process, extensive sub-paragraph marking was carried out.

4.5 Romanian

Contributors: Dan Tufiş and Ştefan Bruda (RACAI), Lidia Diaconu, Călin Diaconu (ICI)

4.5.1 Description of the Corpus

The contents of the Romanian MULTEXT-East newspaper corpus are 128 articles from the daily newspaper "România Liberă", Bucharest, from 12 Apr. 1995.

The digital source used as the basis of encoding was provided, unofficialy by one of the collaborators of "România Liberă" and consisted of data WORD files made by the newspaper's editor.

The Romanian site didn't obtained a license agreement, but a verbal approval for the use of these articles for the purposes of the MULTEXT-East project.

As computed by the Unix program `wc` over the whole CES-1 document, the Romanian newspaper corpus has 26448 words.

4.5.2 Structure of the Corpus

The corpus body consists of 128 <**div type=article**>, each of which contains one file from the original digital data, in most cases constituting one article, and sometimes a series of articles. The articles are grouped by <**div type=page**>.

Each <**div type=article**> begins with one or more <**head**>. Some articles end with <**byline**>.

The text is segmented into paragraphs, with no other paragraph level tagging, except the tags discussed above and <**sp**> for interviews. Sub-paragraph tagging consists of <**hi**> and <**q**>.

The tag usage for the newspaper corpus is shown below.

```
<tagsdecl>
  <tagusage gi=body occurs=1></tagusage>
  <tagusage gi=div occurs=138></tagusage>
  <tagusage gi=head occurs=179></tagusage>
  <tagusage gi=hi occurs=681></tagusage>
  <tagusage gi=p occurs=573></tagusage>
  <tagusage gi=q occurs=76></tagusage>
  <tagusage gi=text occurs=1></tagusage>
  <tagusage gi=byline occurs=77></tagusage>
  <tagusage gi=sp occurs=26></tagusage>
</tagsdecl>
```

Example from the corpus:

```
<div type=page n=1 id="roRL12Apr.1">
<div type=article>
<head>
DUS-&Icirc;NTORS</head>
<head>
Raport pentru stomacuri</head>
<p>
Proverbele sunt un tezaur de &icirc;n&tcedil;elepciune. Nu toate
&icirc;ns&abreve;. Unele
proverbe &scedil;i-au pierdut &icirc;n&tcedil;elesul, dar continu&abreve;
s&abreve; beneficieze
de prestigiul tradi&tcedil;ional al zicerilor populare. Un proverb foarte
```

```
circulat spune c&abreve; <q rend=dblq>Prostul nu doarme de grija
altuia</q>. C&acirc;ndva, &icirc;l denun&tcedil;a pe
b&abreve;g&abreve;re&tcedil;, pe omul care nu-&scedil;i vede de treaba lui.
Cel pu&tcedil;in, a&scedil;a cred, fiindc&abreve; numai un prost poate
s&abreve; cread&abreve; c&abreve; grijile - adic&abreve; temerile, fricile,
nefericirile - altora nu ne privesc. C&abreve; singura solu&tcedil;ie,
c&acirc;nd al&tcedil;ii sunt plini de griji, e s&abreve; dormi bine.
Proverbul acesta func&tcedil;ioneaz&abreve; negativ, ca un tezaur de
suficien&tcedil;&abreve;. Foarte multe proverbe sunt reflexul unor timpuri
dominate de spaim&abreve; &scedil;i resemnare. &Scedil;i,
bine&icirc;n&tcedil;eles, de prostie. De grija altora, cei care nu dorm
sunt &icirc;ntotdeauna de&scedil;tep&tcedil;ii. Pu&tcedil;ini, mul&tcedil;i
c&acirc;&tcedil;i avem. Zicala citat&abreve; con&tcedil;ine punctul de
vedere al omului m&abreve;rginit. E izb&acirc;nda vremelnic&abreve; a <hi
rend=dblq>maselor</hi>. O form&abreve; de cinism mitoc&abreve;nesc devenit
folclor. &Icirc;n tramvaiul 34, un n&abreve;t&abreve;r&abreve;u
f&abreve;r&abreve; griji cugeta la intelectualitate. <q
rend=dblq>A&scedil;a cum m&abreve; vede&tcedil;i</q> - zicea el - <q
rend=dblq>nu m&abreve; dau pe zece profesori</q>. A devenit un obicei ca
dasc&abreve;lul s&abreve; fie unitatea de m&abreve;sur&abreve; a
importan&tcedil;ei altor profesiuni. Suntem, b&abreve;nuiesc, singura
&tcedil;ar&abreve; din lume care scoate anecdote pe seama
&icirc;nv&abreve;&tcedil;&abreve;torilor. Dac&abreve; asta se mai
&icirc;nt&acirc;mpl&abreve; &scedil;i-n alte locuri, situa&tcedil;ia e
grav&abreve;. E ne&icirc;ndoielnic c&abreve; glumele proaste despre
intelighen&tcedil;ie nu le nasc dec&acirc;t cei care, ca &scedil;i
g&acirc;nditorul public din tramvai, se socotesc mai de&scedil;tep&tcedil;i
ca oamenii cu studii. Cum s-a ajuns oare aici? Ceva s-a schimbat &icirc;n
Rom&acirc;nia. Un lucru e ca &icirc;nainte: ordinea social&abreve;.
Intelectualitatea vine tot dup&abreve; clasa muncitoare &scedil;i
&tcedil;&abreve;r&abreve;nimea muncitoare. Nimeni, de la putere, nu a avut
curajul s&abreve;
pun&abreve; treburile la punct. Raporturile puterii cu intelectualitatea
poart&abreve; pecetea
unei ostilit&abreve;&tcedil;i mocnite. Un nenorocit de parlamentar
majoritar spunea, referitor
la exodul creierelor rom&acirc;ne&scedil;ti: <q rend=dblq> Cine vrea
s&abreve; plece, e liber
s&abreve; plece. Nu &tcedil;inem pe nimeni cu sila!</q>. Ca &scedil;i
m&abreve;rginitul din
tramvai, suficientul din parlament se sim&tcedil;ea dezlegat s&abreve; se
exprime a&scedil;a
deoarece, &icirc;n cinci ani, nici pre&scedil;edintele, nici
prim-mini&scedil;trii nu au referit
niciodat&abreve; clar &scedil;i programatic la problemele
intelectualit&abreve;&tcedil;ii.
O guvernare de stomacuri pentru stomacuri, de stomacuri cu somnul gros
&scedil;i ad&acirc;nc.</p>
<byline>
Tudor Octavian</byline>
</div>
.
.
.
```

4.5.3 Structure of the Original

The digital source, used as the basis of encoding, consisted of WORD files, in general, one file per article, made by the newspaper's editor.

4.5.4 Markup Process

Due to the lack of the printed versions, no hilighting marking has been provided, except for the **rend=dblq** marking found in the text. Also, the text contained a number of typo errors, which were also in the printed version. We corrected these errors in the newpaper corpus.

4.6 Slovene

Contributors: Tomaž Erjavec (IJS) and Miro Romih (Amebis d.o.o.)

4.6.1 Description of the Corpus

The contents of the Slovene MULTEXT-East newspaper corpus are 45 articles from the daily newspaper "Dnevnik", Ljubljana, in the period August to October 1995. The digital source used as the basis of encoding was provided, via Amebis d.o.o., by Dnevnik, and consisted of data files made by the newspaper's editor, with idiosyncratic markup and embedded comments for the typesetters. From the large selection of articles originally obtained from Dnevnik, the amount needed by the MULTEXT-East project was selected, choosing longer articles, with an eye towards their being "journalistic" (e.g. discarding serial novels), with a well defined structure.

The Slovene site obtained a license agreement allowing the use of these articles for the purposes of the MULTEXT-East project, signed by the responsible person at Dnevnik.

As computed by the Unix program `wc` over the whole CES-1 document, the Slovene Fiction corpus has 107104 words.

4.6.2 Structure of the Corpus

The corpus body consists of 45 <**div type=article**>, each of which contains one file from the original digital data, in most cases constituting one article, and sometimes a series of articles. In the current version, the article first includes a SGML comment giving the name of the original file. This is followed by a <**opener**>, which contains the <**date**>, giving the date of processing of the file. This date is therefore equal or smaller to the date on which the publication actually appeared.

Each <**div type=article**> contains one or more <**div type=articletext**> containing the actual text(s) of the article; these are optionally followed by one or more <**div type=frame**> or <**div type=figure**>, to be explained below.

The <**div type=articletext**> usually starts off with one or more <**head**> tags, giving the headline(s), and one or more <**div type=articlepart**> giving the sections of the article, each of which can have its own <**head**>.

Articles can also have so called “boxes” or “frames” , i.e. texts given in a frame in the middle of the article, usually passages a few paragraphs long giving background material to the article. These are collected at the end of articles in their own <**div type=frame**>.

Captions to the pictures accompanying the articles were also included in the corpus. Same as frames, they are given in a series at the end of <**div type=article**> and have the following structure:

```
<div type=figure>
<figure>
<head>The text of the figure’s caption</head>
<figdesc>Slika izpuscena (i.e. ‘‘picture omitted’’)</figdesc>
</figure>
</div>
```

The <**div**> elements have the **n** attribute, giving the successive number of the appropriate level of the <**div**>, and the **id** attribute, the value of which has the prefix **dnv** with the laddered section numbers following, separated by periods, e.g. <**div type=articlepart n=6 id=dnv.33.1.6**>.

The frames and figures are problematic for this numbering and id scheme, as they appear at the same level as the article texts — they have been given their own identifiers, **fr** and **fg** respectively. So, for example the first frame of article 8 has the open tag <**div type=frame n=1 id=dnv.fr.8.1**>.

Names of document authors are included where they appeared in the original, usually at the end of the articles; they are marked-up as <**byline**><**docauthor**>Author</**docauthor**></**byline**>.

The text is segmented into paragraphs, with no other paragraph level tagging, except the tags discussed above.

Sub-paragraph tagging consists of <**name**>, and <**q**>. These categories were tagged where they could be automatically inferred from control codes in the original.

No **rend** attributes are included. The quote marks are retained in the <**q**> data, with a double quote being consistently used as the ’top-level’ quote, and apostrophe in possible embedded quotes, usually denoting a “so-called”, which is very popular in the Slovene press.
Example from the corpus:

```
<div type=article n=33 id=dnv.33>
<!-- Original file: 01klolet.t -->
<opener>
<date>10/8/1995</date>
</opener>

<div type=articletext n=1 id=dnv.33.1>
<head>Ali je kaj narobe z varnostjo na"sega zra"cnega prometa</head>
<head>Letala nam padajo z neba</head>

<div type=articlepart n=1 id=dnv.33.1.1>
<head>Letos doslej kar dvanajst nesre"c in vsaj pet incidentov,
"zivljenje so izgubili trije ljudje - Vzrokov je ve"c, v"casih tudi
lahkomiselnost</head>
```

```
<p>
Pilot in kopilot zgorela, Zasilni pristanek se ni posre"cil, "crna
nedelja v zraku in na tleh, MAG na hrbtu, Pilot in dekle "cude"zno
```

4.6.3 Structure of the Original

The digital source used as the basis of encoding was provided, via Amebis d.o.o., by Dnevnik. It consisted of data files, one file per article, made by the newspaper's editor. Each file starts off by header information, followed by the text. The text itself contains idiosyncratic markup and embedded comments for the typesetters.

```
>beginh<
user=Tone
status=normal
format=dnevnik
format=nd
creator=Tone,4/8/1995,10:26:6
editor=Tone,4/8/1995,10:26:6
editor=Sonja,4/8/1995,12:24:56
editor=Tone,10/8/1995,15:49:4
abstract=0
-so

Ali je kaj narobe z varnostjo naega zra\248nega prometa
Letala nam padajo z neba
Letos

h=904.96dd
version=16
>endh<
-so

>12n<Ali je kaj narobe z varnostjo naega zra\248nega prometa

>54a<Letala nam padajo z $
neba

>14pn<Letos doslej kar dvanajst nesre\248 in vsaj pet incidentov,
\236ivljenje so izgubili trije $
ljudje - Vzrokov je ve\248, v\248asih tudi lahkomiselnost

Pilot in kopilot zgorela, Zasilni pristanek se ni posre\248il, \252rna
nedelja v zraku in na tleh, MAG na hrbtu, Pilot in dekle \248ude\237no $
pre\236ivela, Trd pristanek lahkega letala v koruzi, Helikopter po
nesre\248i pobegnil>tp< To je le nekaj \248asopisnih naslovov zadnjih dveh $
```

4.6.4 Markup Process

The original digital (and only) source were files made by the newspaper's editor, with idiosyncratic markup and embedded comments for the typesetters. From the large selection of articles originally obtained from Dnevnik, the amount needed by the MULTEXT-East project was selected. After discussions on the desired CES1 structure, files were automatically converted, by Amebis d.o.o., into CES1 marked-up files, and at the same time corrected for typos with a spelling-checker. This version has been then additionally hand-corrected and the header made to produce a standalone CES1 SGML document.

Chapter 5

Multilingual Parallel Speech Corpus

For the MULTEXT-East speech corpus, a small sample of the English part of the EUROM1 multilingual speech database was selected. This text was translated into the six languages and, except for Bulgarian and Czech, recorded for one male speaker and digitized in accordance with the EUROM recommendations. The seven texts were, in addition to the EUROM format, also encoded in CES, as **cesDoc** elements.

The texts chosen comprise 40 passages (4 x 10, designated blocks: `O`, `P`, `Q`, `R` / `0`–`9`) of 5 thematically linked sentences. As an example we give below the English blocks `O0` and `O1`:

```
                         BLOCK: O0

Last week my friend had to go to the doctors to have some injections. She
is going to the Far East for a holiday and she needs to have an injection
against cholera, typhoid fever, hepatitis A, polio and tetanus. I think she
will feel quite ill after all those. She is going to get them all done at
once, at one session. I shan't feel sorry for her though!


                         BLOCK: O1

I have a problem with my water softener.  The water-level is too high and
the overflow keeps dripping. Could you arrange to send an engineer on
Tuesday morning please? It's the only day I can manage this week. I'd be
grateful if you could confirm the arrangement in writing.
```

The translations into the MULTEXT-East languages were 'localised', i.e. the situations described in the monologues were translated as if the speaker were describing a situation in his native land. For example, local place names were used instead of the British ones.

The text were spoken by one native male speaker. The recording was performed as close as possible to the EUROM guidelines, which are as follows:

- B&K 4165 condenser measurement microphone (or other high quality microphone) positioned for zero degree incidence, 50cm from lips and 15 degrees off the straight ahead position from the head.

- The recording procedure done in an anechoic room or a recording studio.

The digitisation features of the recordings were the following:

- Sampling frequency: 20000 Hz.
- Samples coded on 16 bits, linear, twos-complement quantisation.

5.1 Organisation of the MULTEXT-East EUROM Corpus

In the corpus there are two types of files:

- *speech files*: digitised speech samples in a simple binary file (NO header);
- each speech file is associated with an ASCII *description file*, this association relaying on filenames.

All the filenames have the structure: `name.suffix`

The *name* is composed of:

1. the first 2 characters are the initials of the speaker.
2. the 2 following characters are the nomenclature (number) of the passage.
3. the 4 following figures are free

The speech sample file suffix is, for example, `.pss` (`p` = passage; `s` = Slovene; `s` = speech sample).

The following one letter codes were used for the MULTEXT-East languages:

- `e` = English
- `b` = Bulgarian
- `c` = Czech
- `t` = Estonian
- `h` = Hungarian
- `r` = Romanian
- `v` = Slovene

The description file *suffix* is `pXo` with:

- `p` = passage

- `X` = the initial of the country name
- `o` = associated description file

Here is an example of an English associated file (ie suffix `.peo`):

```
------------------------------------------------------------------------
HD: V3.0
TYP: orthographic
DBN: EUROM_1
VOL:
DIR:
SRC: FA000079.PES
TXF: 00.TXT
CMT: Information about the recording session
SAM: 20000
BEG: 0
END: 406271
RED: 18/Jan/91
RET: 14:15:50
REP: UCL
SNB: 2
SBF: 01
SSB: 16
RCC: 2
NCH: 2
SPI: M, 48, BRITISH
PCF: PASSAGE.DES
CMT: Information about the labelling session
EXP:
SYS:
DAT:
SPA:
CMT: Item: label start, end, input gain, min level, max level, string
LBD:
LBR: 0, 406271, 6, -3706, 4534, Last week my friend had to go to the doctors to
EXT: have some injections. She  is going to the Far East for a holiday and she
EXT: needs to have an injection against cholera, typhoid fever, hepatitis A,
EXT: polio and tetanus. I think she will feel quite ill after all those. She is
EXT: going to get them all done at once, at one session. I shan't feel sorry
EXT: for her though!
LB2: 0, 406271, 0, -6143, 9339
ELF:
---------------------------------------------------------------------------
```

The following are the minimal fields that must be present in a description file:

`TYP, DBN, SRC, TXF, SAM, SNB, SBF, SSB, BEG, END, RED, SPI, PCF,`
`LBR and following EXT`

5.2 Structure of the CES Speech Corpus

In addition to the EUROM format, the textual part of the MULTEXT-East speech corpus was also encoded in CES. For each of the seven languages, the passages were encoded as a **cesDoc** element, with a header describing the corpus component, and the **text** containing the passages. Each passage was encoded as a **div**ision of type **block**, containing a **head** and one paragraph. The text itself was sentence segmented and IDed. The following English `BLOCK 00` illustrates the resulting structure:

```
<div id="Sen.1" type="block" n="00">
<head>*BLOCK: 00</head>
<p id="Sen.1.2">
<s id="Sen.1.2.1">Last week my friend had to go to the doctors to have
some injections.</s>
<s id="Sen.1.2.2">She is going to the Far East for a holiday and she
needs to have an injection against cholera, typhoid fever, hepatitis
A, polio and tetanus.</s>
<s id="Sen.1.2.3">I think she will feel quite ill after all those.</s>
<s id="Sen.1.2.4">She is going to get them all done at once, at one
session.</s>
<s id="Sen.1.2.5">I shan't feel sorry for her though!</s>
</p>
</div>
```

The complete CES encoded MULTEXT-East speech corpus has the following tag usage:

- **text** = 7, **body** = 7
- **div** = 280, **head** = 280, **p** = 280
- **s** = 1377

It is interesting to note the number of sentences; the EUROM states that the passages consist of "5 thematically linked sentences". Thus the number of all the sentences should be be 1400 (200 x 5), but is actually less. This holds even for the English original, which has 196 sentences instead of 200.

Appendix A

TELRI Appendix 1: Additional "1984"

Workpackage Coordinator: Tomaž Erjavec
Contributors:
Latvian: Andrejs Spektors
Lithuanian: Andrius Utka
Serbo-Croat: Cvetana Krstev, Duško Vitas

This report is an appendix to the MTE D2.1 F report, which documents the Deliverable D2.1 carried out within the framework of the Copernicus 106 MULTEXT-East (Multilingual Text Tools and Corpora for Eastern and Central European Languages) Project. This deliverable consisted of collecting and annotating a multilingual corpus, totaling cca 2 million words. The central component of this corpus is the parallel corpus, composed of Orwell's novel "1984" in the English original, along with translations into Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene. In addition to basic structural encoding, this corpus was sentence aligned between the translations and English, and each language component tagged for part-of-speech.

Due to the Copernicus concerted action TELRI, (Trans-European Language Resources Infrastructure), it was possible to add three new translations to the existing seven-language "1984" corpus. This Appendix documents these new "1984" language components, namely the Latvian, Lithuanian, and Serbo-Croat translations of "1984".

In order to make this Appendix as self-contained as possible, we give in the following sections the details on the encoding of the TELRI "1984" additions as a whole, highlighting the points common both the MULTEXT-East and the TELRI "1984" corpus: these sections are very similar to the Overview section of the "1984" corpus collection report, i.e. MULTEXT-East D2.1 F, Section 2.1.

These overview sections are followed by sections detailing the encoding of the TELRI translations. For each language we give a description of the corpus, its structure, the structure of the original from which the CES version was derived, and the markup process that led from one to the other.

Note that the added alignment markup of the "1984" TELRI corpus, which is contained in separate SGML documents, hyperlinked to the "1984" corpus, is in structure identical to the alignments of the MULTEXT-East "1984" corpus. The reader is referred to the MULTEXT-East Deliverable D2.3 F for the description of these alignments.

A.1 Overview

The encoding of the new translations was harmonised, as much was possible, with the MULTEXT-East "1984" corpus components. Due to tight time and effort constraints, the encoding is often less detailed, and there might be more encoding errors than in the MULTEXT-East "1984". In general, however, the encodings are more similar than different, esp. as regards their segmentation, so as to simplify sentence alignment.

Below we give an estimate for the number of words, by language. The wordcounts were produced by removing the SGML tags from the texts and then using a 'wc'-like procedure. The translations described here are in boldface:

English	104.302
Romanian	101.460
Slovene	91.619
Serbo-Croatian	**89.749**
Bulgarian	87.235
Latvian	**81.956**
Czech	80.366
Hungarian	81.147
Estonian	79.334
Lithuanian	**71.252**

Wordcounts in Orwell's "1984"

A.2 Corpus Encoding

The "1984" corpus has been encoded in SGML, in the Corpus Encoding Specification (CES) Document Type Declaration. The CES DTD, along with documentation can be obtained from `http://www.cs.vassar.edu/CES/`.

The complete MULTEXT-East CES corpus is encoded as a <**cesCorpus**> document, which comprises a header and the 26 component parts (seven for "1984") of the corpus, each encoded as a <**cesDoc**> element. The TELRI "1984" translations have not been included in this <**cesCorpus**>, but are encoded as separate <**cesDoc**> SGML documents, with no common header.

Each <**cesDoc**> document is stored in a separate file. All system identifiers (i.e. filenames) are encapsulated in the catalog, which is structured according to the SGML Open Technical Resolution 9401:1997. There the TELRI "1984" corpus components are given the following PUBLIC identifiers:

```
-//TELRI//DOCUMENT CES1 1984//LV
-//TELRI//DOCUMENT CES1 1984//LT
-//TELRI//DOCUMENT CES1 1984//SH
```

Each <**cesDoc**> element is also marked with the **lang** attribute that gives the language of the corpus component (but note that all <**cesHeader**> elements are marked as English). For language IDs, the two letter ISO 639 values have been used. The definitions for the CES **language** elements are encapsulated in the file with the PUBLIC identifier:

```
ISO 639-1988//ENTITIES Languages//EN
```

These entities encompass the European languages; we give here the definitions for the three TELRI languages:

```
<language id=lv iso639=lv>Latvian;Lettish</language><!--BALTIC-->
<language id=lt iso639=lt>Lithuanian</language>     <!--BALTIC-->
<language id=sh iso639=sh>Serbo-Croatian</language> <!--SLAVIC-->
```

For (language specific) character representation, the documents use SGML entities from the following entity sets:

```
ISO 8879-1986//ENTITIES Added Latin 1//EN
ISO 8879-1986//ENTITIES Added Latin 2//EN
ISO 8879-1986//ENTITIES Russian Cyrillic//EN
ISO 8879-1986//ENTITIES Non Russian Cyrillic//EN
```

The first two are used for Latvian in Lithuanian, while Serbo-Croatian (indirectly) uses all four. Namely, the Serbo-Croatian translation is from Serbia, where text can be written either in the Latin or in the Cyrillic alphabet. Therefore the entities that are used in the Serbo-Croatian “1984” provide mapping to either Added Latin or to Cyrillic entities.

For each language (the reference to) its language specific character set entities has been encapsulated in files with the following PUBLIC identifiers:

```
-//MTE//ENTITIES Latvian//EN
-//MTE//ENTITIES Lithuanian//EN
-//MTE//ENTITIES Serbian//EN
```

It should be noted that, in the current version, only the complete MULTEXT-East `-//MTE//DOCUMENT CES1//` corpus constitutes a valid SGML document. Conversely, each TELRI “1984” component is a valid document, i.e. contains the SGML prolog.

To retain (future) compatibility with the MULTEXT-East “1984”, the structure of the TELRI “1984” is slightly more complicated than would be necessary for ’stand alone’ documents. As an example we give the prolog of `-//TELRI//DOCUMENT CES1 1984//LT`:

```
<!DOCTYPE cesDoc PUBLIC "-//CES//DTD cesDoc//EN" [
  <!ENTITY % ONECOMPONENT "INCLUDE">
  <!ENTITY ISOlang PUBLIC "ISO 639-1988//ENTITIES Languages//EN">
  <!ENTITY % MTElt PUBLIC "-//MTE//ENTITIES Lithuanian//EN">
  %MTElt;
]>
```

Marked sections have been made use of: if `ONECOMPONENT` is set to `INCLUDE`, then <**language**> definitions are included in the <**cesDoc**> header. In the (MULTEXT-East) corpus, ONECOMPONENT is set to IGNORE, and the the <**language**> definitions are a part of the <**cesCorpus**> header.

All the corpus components have been encoded at least up to CES level 1. Level 1 markup includes a TEI-like header (file, encoding, profile and revision descriptions), and universal text elements down to the level of the paragraph, e.g. textual divisions, paragraphs, titles and headings, footnotes, tables and poems. Some CES 2 level markup has also been included, e.g. quoted material (<**quote**>, <**q**>), and rendition information. Finally, some CES 3 level markup is also present, namely sentence markup with <**s**>.

To illustrate the kinds of elements used in the “1984” corpus, and to show their distributions across the language components we give the following table, which gives the contents of the <**tagUsage**> elements in the headers of the respective language components. For comparative purposes, we give the numbers for the complete “1984” corpus:

GI	EN	BG	CS	ET	HU	RO	SL	LV	LT	SH
abbr	38	28	23	73	38	3	26			7
date	40	40	39	18	39		33			39
body	1	1	1	1	1	1	1	1	1	1
div	28	28	28	28	28	28	28	28	28	28
foreign	39	29	91	93	43	430	7			7
head	1	1	1	5	5	28	29	28	28	27
hi	103	103	75	183	71	413	242	129	136	323
item	4	4	4	4	4	4	4	4	4	4
l	32	26	33	32	32	26	34	32	36	32
list	1	1	1	1	1	1	1	1	1	1
mentioned	261	256	244	44	281					
name	1744	1704	2181	2457	1843	2157	1327	163		1371
note	2	8	2	2	2	3	1	1	1	1
num	52	34	48	14	10			1		
p	1286	1321	1285	1289	1292	1335	1288	1332	1331	1279
poem	10	7	11	10	10	7	10	10	12	10
ptr	2	8	1	2	2		1	1	1	
q	2209	1203	2208	2192	2197	2137	2260	210	40	2246
quote	35	34	36	35	35	23	35	25	38	35
s	6701	6649	6714	6658	6732	6487	6689	6690	6675	6652
title	46	41	45	29	40	1	10	1	7	4
term			2							

Tag usage in Orwell’s “1984”

For the purpose of alignment it was important to ensure that the gross structure of all the languages was as similar as possible. The translations were encoded taking the English digital version as as the norm, with errors of alignment guiding the harmonisation. The elements that have been harmonised can be seen in the table above, as they occur the same number of times in all the languages. In particular, each of the “1984”s has one <**body**>, which consists of three

<div type=part n=1, 2, 3> and of one **<div type=part n=appendix>**. Each part, except the Appendix, is further subdivided into a number of **<div type=chapter n=1, 2, . . . >**.

A.2.1 Sentence segmentation

Prior to sentence alignment, the "1984" had to be sentence segmented. In inserting the **<s>** markup, the well known problem of crossing **<q>** and **<s>** hierarchies was avoided by not marking up (problematic) **<q>** elements: these are indicated by the original rendering. In the cases where **<q>** elements were already in place, and crossed with the **<s>** structure, the **<q>** elements were split and marked by **type=MI** (Machine Inserted).

The automatically inserted **<s>** elements were then hand validated via the process of alignment: where the S alignment between the English version and a translation was not one-to-one, there was a fair chance of it being not a difference in translation, but of a wrongly placed **<s>**. Such alignments were checked, and the **<s>** tags corrected where necessary.

A.2.2 ID marking

The `ider` program, written by Greg Priest-Dorman, was used to automatically assign unique identifiers to the following structural elements in the "1984" corpus:

- **<div>**
- **<p>**, **<list>**, **<poem>**
- **<s>**, **<item>**, **<l>**

The values of the **id** attribute are strings composed of the initial letter **O** identifying the **O**rwell corpus, the two letter ISO 639 code identifying the language component of the corpus (`en, bg cs, et, hu, ro, sl`), followed by numbers separated by periods; these give the position of the tag in the SGML tree of the document, rooted at the BODY element. The following example illustrates the ID naming scheme:

```
  <body lang="sl" id="Osl">
    <div id="Osl.1"  type=part n=1>
    <head>Prvi del</head>
      <div id="Osl.1.2" type=chapter n=1>
        <head>I</head>
        <p id="Osl.1.2.2">
          <s id="Osl.1.2.2.1">
            Bil je jasen, mrzel aprilski dan in ure so bile trinajst.
          </s>
...
        <quote id="Osl.1.2.17" rend="CN IT"><p id="Osl.1.2.17.1">
            <s id="Osl.1.2.17.1.1">
              <date iso8601="1984-04-04">4. april 1984</date>
            </s>
          </p></quote>
...
```

```
    <quote id="Osl.1.8.18" rend="CN IT"><poem id="Osl.1.8.18.1">
        <l id="Osl.1.8.18.1.1">Tam pod kostanjevim drevesom</l>
        <l id="Osl.1.8.18.1.2">izdala si me,</l>
        <l id="Osl.1.8.18.1.3">izdal sem te,</l>
        <l id="Osl.1.8.18.1.4">ne da bi trenila z o&ccaron;esom.</l>
      </poem></quote>
```

A.2.3 HTML rendering

For the corpus components, HTML translations of <**cesDoc**> headers and samples of the texts are also available. These were obtained with custom-written CES 2 HTML translation maps, using the Fred software package (see `http://www.oclc.org/fred/`). The mapping does not preserve information explicitly, but 'renders' it. For example, <**name**> is represented as <**b**> and <**respType**> as '(...)'.

A.3 Latvian

Contributors: Andrejs Spektors and Tomaž Erjavec

A.3.1 Description of the Corpus

The digital source of the Latvian version of "1984" was created at the Artificial Intelligence Laboratory, University of Latvia, Riga. The publication and translation in Latvian that was used for the creation of the digital source is the one by Avots, Riga, published in 1990. The digital source was obtained by OCR from the printed edition.

The copyright of the translation of the Latvian "1984" is held by the Latvian copyright agency "AKKA/LAA". We currently do not have permission to distribute this translation further.

The Latvian version of "1984" contains 81.956 words, as indicated in the header of the tagged version.

A.3.2 Structure of the Corpus

The Latvian "1984" corpus body consists of three <**div type=part n=1, 2, 3**> and of one <**div type=part n=appendix**>. Each part is further subdivided into a number of <**div type=chapter n=1, 2, . . .** >. In the Latvian version, each <**div**> is followed by a <**head**>, giving the part or chapter number. Counting of chapters starts from **1** in every part.

Elements <**body**>, <**div**>, <**head**>, <**item**>, <**l**>, <**list**>, <**note**>, <**p**>, <**poem**>, <**ptr**>, <**quote**>, <**text**> are used so that to be in harmony with the English 1984 for MULTEXT-East (`-//MTE//TEXT CES1 1984//EN`); the differences are due only to the differences between the English electronic and Latvian printed version.

Of the possible sub-paragraph tags, only <**hi**>, <**mentioned**>, <**name**>, <**q**>, and <**title**> are used. The <**q**> tag is used only to denote quoted words, but not direct speech. Rendering information is given as the CES conformant two-letter value of the **rend** attribute. It has been in most cases included with the appropriate tags. The values for **rend** attribute are: CA, CE, and IT.

The mark-up is even in all chapters of the novel, i.e. no chapters could be distinguished as having more **type** information on tags than others.

The text has been automatically sentence segmented, and the segmentation hand-validated. The <**body**>, <**div**>, <**p**>, <**poem**>, <**list**>, <**l**>, <**item**>, and <**s**> tags have been marked with the **id** attribute.

Page brakes from the original have been preserves as comments.

The following is an example from the Latvian "1984" corpus:

```
<p id="Olv.1.2.9"><s id="Olv.1.2.9.1">Patie&imacr;bas ministrija
&mdash; jaunrun&amacr; to sauca par Patminu &mdash; krasi
at&scaron;&kcedil;&imacr;r&amacr;s no visiem citiem redzamiem
objektiem.</s>
<s id="Olv.1.2.9.2">T&amacr; bija milz&imacr;ga, piram&imacr;dai
l&imacr;dz&imacr;ga mirdzo&scaron;i balta betona celtne, kas
daudz&amacr;m citu par citu augst&amacr;k&amacr;m teras&emacr;m
sl&emacr;j&amacr;s debes&imacr;s tr&imacr;ssimt metrus.</s>
<s id="Olv.1.2.9.3">No vietas, kur Vinstons st&amacr;v&emacr;ja,
var&emacr;ja salas&imacr;t t&amacr;s baltaj&amacr; fas&amacr;d&emacr;
skaistiem burtiem iecirstos tr&imacr;s partijas sauk&lcedil;us:</s>
<q rend="CE CA" type=slogan>KAR&Scaron; IR MIERS</q>
<q rend="CE CA" type=slogan>BR&Imacr;V&Imacr;BA IR VERDZ&Imacr;BA</q>
<q rend="CE CA" type=slogan>NEZIN&Amacr;&Scaron;ANA IR SP&Emacr;KS</q>
</p>
```

A.3.3 Structure of the Original

The electronic version that was obtained by OCR preserves most of the visual layout peculiarities in electronic form that are in the printed version of "1984" including line breaks, end-of-line hypenation, centering, capitalization, italic print, page numbers, Latvian characters and foreign characters.

A.3.4 Markup Process

The electronic version of "1984" that was created from the printed version by OCR was taken as the basis for the encoding. The typical OCR mistakes were corrected manually. As this version has similar layout to the printed version of "1984" it enabled to mark up many visual distinctions semi-automatically. It was proofread and marked up to CES1 conformance. In the process, some typographical errors were discovered not only in the digital version, but also in the printed edition of the Latvian translation of "1984". The transliteration of Latvian characters by SGML entities was performed automatically. A number inconsistencies and anomalies were discovered and corrected in the process of aligning Latvian translation to the original.

A.4 Lithuanian

Contributors: Andrius Utka and Tomaž Erjavec

A.4.1 Description of the Corpus

The digital source of the Lithuanian version of "1984" was created at the Centre of Computational Linguistics of the Kaunas Vytautas Magnus University. The publication and translation in Lithuanian that was used for the creation of the digital source is the one by "Vyturys", Vilnius, published in 1991. The digital source was obtained by OCR from the printed edition.

The digital source for the Lithuanian "1984" has the same distribution restrictions as those imposed by the MULTEXT-East project, namely, that the use of the materials is free for academic purposes.

We currently do not have written permission to use the text, however, we do have verbal assurance from the translator that it is acceptable to use the Lithuanian translation of "1984" for the purposes of the TELRIproject.

The Lithuanian version of "1984" contains 71.210 words, as indicated in the header of the tagged version.

A.4.2 Structure of the Corpus

The Lithuanian "1984" corpus body consists of three <**div type=part n=1, 2, 3**> and of one <**div type=part n=appendix**>. Each part is further subdivided into a number of <**div type=chapter n=1, 2, . . .** >. In the Lithuanian version, each <**div**> is followed by a <**head**>, giving the part or chapter number. Counting of chapters starts from **1** in every part.

Elements <**body**>, <**div**>, <**head**>, <**item**>, <**l**>, <**list**>, <**note**>, <**p**>, <**poem**>, <**ptr**>, <**quote**>, <**text**> are used so that to be in harmony with the English 1984 for MULTEXT-East (`-//MTE//TEXT CES1 1984//EN`); the differences are due only to the differences between the English electronic and Lithuanian printed version.

Sub-paragraph tags, i.e. <**abbr**>, <**date**>, <**foreign**>, <**mentioned**>, <**name**>, <**num**> are not used, except for <**hi**>,<**q**>,<**title**> that are used inoconsistently. Rendering information is given as the CES conformant two-letter value of the **rend** attribute. It has been in most cases included with the appropriate tags, except for the default preceding `—` of the <**q**> tag. The values for **rend** attribute are: CA, CE, CN, IT, and '*'.

The mark-up is even in all chapters of the novel, i.e. no chapters could be distinguished as having more **type** information on tags than others.

The text has been automatically sentence segmented, and the segmentation hand-validated. The <**body**>, <**div**>, <**quote**>, <**p**>, <**poem**>, <**list**>, <**l**>, <**item**>, <**s**>, and <**q**> tags have been marked with the **id** attribute.

The following is an example from the Lithuanian "1984" corpus:

```
<p id="Olt.1.2.8">
<s id="Olt.1.2.8.1">Teisyb&edot;s ministerija&mdash;naujakalbe
<ptr id="Olt.1.2.8.1.1" n="1" target=N1 rend=asterisk target= "Olt.1.2.11">
Teisybmin&edot; &mdash; akivaizd&zcaron;iai skyr&edot;si nuo vis&uogon; kit&uogon;
aplinkini&uogon; pastat&uogon;.</s><s id="Olt.1.2.8.2">Tai buvo giganti&scaron;kas
piramid&edot;s pavidalo statinys, &zcaron;vilgantis baltu betonu, terasa po terasos kylantis &iogon;
300 metr&uogon; auk&scaron;t&iogon;.</s><s id="Olt.1.2.8.3">Vinstonas i&scaron; &ccaron;ia dar
gal&edot;jo perskaityti grak&scaron;&ccaron;iomis raid&edot;mis baltame fone
&scaron;vie&ccaron;ian&ccaron;ius tris partijos &scaron;&umacr;kius:
```

```
<q id="Olt.1.2.8.3.1" rend="CE CA" type=slogan>KARAS&mdash;TAI TAIKA</q>
<q id="Olt.1.2.8.3.2" rend="CE CA" type=slogan>LAISV&Edot;&mdash;TAI VERGIJA</q>
<q id="Olt.1.2.8.3.3" rend="CE CA" type=slogan>NE&Zcaron;INOMAS&mdash;TAI
J&Edot;GA</q>
</s></p>
```

A.4.3 Structure of the Original

The electronic version that was obtained by OCR preserves all the visual layout peculiarities in electronic form that are in the printed version of "1984" including new line, centering, capitalization, bold print, italic print, page numbers, Lithuanian characters and foreign characters.

A.4.4 Markup Process

The electronic version of "1984" that was created from the printed version by OCR was taken as the basis for the encoding. The typical OCR mistakes were corrected manually as well as by spelling checker. As this version has similar layout to the printed version of "1984" it enabled to mark up many visual distinctions semi-automatically. It was proofread and marked up to CES1 conformance. In the process, some typographical errors were discovered not only in the digital version, but also in the printed edition of the Lithuanian translation of "1984". The transliteration of Lithuanian characters by SGML entities was performed automatically. A number inconsistencies and anomalies were discovered and corrected in the process of aligning Lithuanian translation to the original.

A.5 Serbo-Croatian

Contributors: Cvetana Krstev, Duško Vitas, and Tomaž Erjavec

A.5.1 Description of the Corpus

The digital source of the Serbo-Croatian version of "1984" was obtained by Oxford Text Archive. It was prepared, together with an English, Slovenian and Croatian version by the OUCS KDEM service in 1985 for Dr David C Bennett of the School of Oriental and African Studies at London University. The OTA has copies of letters from all the publishers concerned which appear to indicate that academic use of the text is permitted. This version is in OTA labeled as "Serbian" although the imprint of the first edition published by "JUGOSLAVIJA" states that it is "the first edition of '1984' in Serbo-Croatian".

Publication used for the digital source is the second edition published in 1984 by "Beogradski izdavač ko-grafič ki zavod", Belgrade and translated by Vlada Stojiljković. The same translation had two previous editions by "JUGOSLAVIJA" in 1968 and 1977. The OTA version was obtained by OCR from the printed edition.

The digital source for the Serbo-Croatian "1984" has the same distribution restrictions as those imposed by the MULTEXT-East project, namely, that the use of the materials is free for academic purposes.

We currently do not have written permission to use the text, however, we do have verbal assurance from OTA that it is acceptable to use the Serbo-Croatian translation of "1984" for the purposes of the TELRI project.

The Serbo-Croatian version of "1984" contains 89749 words, as indicated in the header of the tagged version.

A.5.2 Structure of the Corpus

The Serbo-Croatian "1984" corpus body consists of three <**div type=part n=1, 2, 3**> and of one <**div type=part n=appendix**>. Each part is further subdivided into a number of <**div type=chapter n=1, 2, . . .** >. In the Serbo-Croatian version, each <**div**> is followed by a <**head**>, giving the part or chapter number. Counting of chapters starts from **1** in every part.

Elements <**body**>, <**div**>, <**head**>, <**item**>, <**l**>, <**list**>, <**note**>, <**p**>, <**poem**>, <**ptr**>, <**quote**>, <**text**> are used so that to be in harmony with the English 1984 for MULTEXT-East (`-//MTE//TEXT CES1 1984//EN`); the differences are due only to the differences between the English electronic and Serbo-Croatian printed version.

Sub-paragraph tags, i.e. <**abbr**>, <**date**>, <**foreign**>, <**mentioned**>, <**name**>, <**num**>, <**hi**>, <**q**>, <**title**> are used extensively, but it is not ceratin that they are not missing occasionally. For example, names have been tagged automatically, and the tagged names hand-validated; therefor all the tagged names are correct, but the text might also contain untagged names. The name tags do not contain the type attribute. In accordance with CES, only proper nouns have been tagged, while adjectives derived from proer nouns, e.g. *Vinstonov*, have not.

Rendering information is given as the CES conformant two-letter value of the **rend** attribute. It has been in most cases included with the appropriated tags. The values for **rend** attribute are: CA, CE, CN, IT, and '*'.

Rendering, including "all caps" capitalisation, has been removed from the tag content.

The mark-up is even in all chapters of the novel, i.e. no chapters could be distinguished as having more **type** information on tags than others.

The text has been automatically sentence segmented, and the segmentation hand-validated. Some <**q**> elements have been split in this process; these are marked with **type=MI**, for "machine inserted". The <**body**>, <**div**>, <**quote**>, <**p**>, <**poem**>, <**list**>, <**l**>, <**item**>, <**s**>, and <**q**> tags have been marked with the **id** attribute.

The following is an example from the Serbo-Croatian "1984" corpus:

```
<p id="Oshs.1.2.8"><s id="Oshs.1.2.8.1">Ministarstvo istine &mdash; u
<name>Novogovoru</name>: <name>Ministin </name> &mdash;
o&sx;tro se razlikovalo od svega ostalog na vidiku.</s>
<s id="Oshs.1.2.8.2">To je
bila ogromna piramidalna gra&dx;evina od svetlucavo
belog betona koja se uzdizala, terasa za terasom, tri
stotine metara u nebo.</s>
<s id="Oshs.1.2.8.3">Sa mesta na kome je <name>Vinston</name>
stajao mogle su se tek razabrati, ispisane elegantnim
slovima na belom zidu, tri parole Partije:
<q rend="CN CA" type=slogan>Rat je mir</q>
<q rend="CN CA" type=slogan>Sloboda je ropstvo</q>
```

```
<!-- pb n=7 -->
<q rend="CN CA" type=slogan>Nezna&nx;e je mocx;</q></s>
</p>
```

A.5.3 Structure of the Original

The electronic version that was obtained by OCR preserves all the visual layout peculiarities in electronic form that are in the printed version of "1984" including new line, centering, capitalization, bold print, italic print, page numbers. Serbo-Croatian characters where encoded as cy, sy, dj, lj, zy, nj, cj, dzy for č, š, đlj, ž, nj, ć, and dž respectively. Two pair of codes were used for each of these upper-case letters: for instance, for Č both CY and Cy were used. Encodings for đ and dž caused problems in quite a number of words: *odjednom*, *podjednako*, *odjek*, *odjeknuti*, *nadjačati*, *odživeti*, etc.

The following is an example from the OTA version:

```
   Ministarstvo istine - u Novogovoru: Ministin -
osytro se razlikovalo od svega ostalog na vidiku. To je
bila ogromna piramidalna gradjevina od svetlucavo
belog betona koja se uzdizala, terasa za terasom, tri
stotine metara u nebo. Sa mesta na kome je Vinston
stajao mogle su se tek razabrati, ispisane elegantnim
slovima na belom zidu, tri parole Partije:

   <1RAT JE MIR>1
   <1SLOBODA JE ROPSTVO>1




                                             [7]
   <1NEZNANJE JE MOCJ>1
```

A.5.4 Markup Process

The OTA version of "1984" was taken as the basis for the encoding. The whole Serbo-Croatian "1984" CES1 corpus was cross-cheked with the printed edition, and the printed edition was used to insert additional (e.g. <**hi**>) markup. In the process, a number of typographical erros were discovered not only in the OTA edition, but also in the printed edition of the Serbo-Croatian translation of "1984". The transliteration of the OTA encoding of Serbo-Croatian characters cy, sy, dj, lj, zy, nj, cj, dzy by SGML entities &cy;, &sx;, &dx;, &lx;, &zx;, &nx;, &cx;, and &dy; was performed automatically and the replacements for dj and dzy were hand-validated. The replacement text for the SGML entities &cy;, &sx;, &dx;, &lx;, &zx;, &nx;, &cx;, &dy; is a character string (nj or lj), a reference of a SGML entity from the standard entity set ISOlat2 (for instance, č) or a combination of those (dž) as Latin alphabet was used for the prined edition.