Corpus Markup

COP Project 106 MULTEXT-East

Work Package WP2 – Task 2.3
Deliverable D2.3 F

Final Report

21 December 1997

Workpackage Coordinators:
Greg-Priest Dorman, Tomaž Erjavec, Nancy Ide, Vladimír Petkevič

Contributors:
English: N.Ide and G.Priest-Dorman
Bulgarian: L.Dimitrova and L.Sinapova
Czech: V.Petkevič, J.Hajič, B.Hladká
Estonian: H.J.Kaalep and L.Paldre
Hungarian: C.Oravecz
Romanian: Ş.Bruda and D.Tufiş
Slovene: T.Erjavec, A.Bizjak, P.Jakopin

Contents

Chapter 1

Introduction

This report documents Deliverable D2.3 carried out within the framework of the Copernicus 106 MULTEXT-East (Multilingual Text Tools and Corpora for Eastern and Central European Languages) Project. The MULTEXT-East task 2.3 consists of added linguistic annotations to the parallel part of the MULTEXT-East multilingual corpus collection. The parallel corpus consists of the novel 1984 by G.Orwell in the seven languages of the project.

The markup of the complete MULTEXT-East corpus to CES Level 1 (header & basic structure markup and sentence segmentation) is given in the D2.1 report "Sample Corpus Collection and Preparation". The Level 1 markup includes a TEI conformant 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, rendition information, and, to varying degrees, abbreviations, dates, names, and numbers. Finally, the parallel "1984" corpus has been sentence segmented, the segmentation manually validated and the structural elements of the corpus marked with unique identifiers.

The present report details the additional markup performed on the "1984" MULTEXT-East corpus, which consists of:

- sentence alignment;
- morphosyntactic tagging.

This additional markup was not encoded in the primary data but, in line with the MULTEXT philosophy, stored in separate SGML documents, and hyperlinked to the primary "1984" data. These documents are also 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 rest of this report explains the structure of the above two annotations to the parallel corpus.

Chapter 2

Sentence Alignment

Each of the six translations of 1984 has been S aligned with the English original, and the alignments hand validated. The alignment is not hierarchical, i.e. division and paragraph level alignments have not been retained, although they have been used in the process of alignment. The S-level elements that have been aligned are the following:

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

The initial S alignment was performed automatically, where different aligners were used for different languages:

- Estonian used the MULTEXT `MtAlign` program by ISSCO;
- Romanian used the NANCY aligner by INRIA/CNRS Nancy;
- the other languages used the 'Vanilla' aligner, written by D. Ridings and P. Danielsson from Göteborg University.

With Vanilla (see *http://svenska.gu.se/PEDANT/workshop/workshop.html*), the texts were first aligned to the paragraph level, these alignments checked and, where necessary, corrected. Once this alignment was correct, the paragraph level links were taken as 'hard' links, and S-level alignment performed. This was again hand-validated, where, in addition to alignment errors, this validation often exposed errors of sentence segmentation.

Automatic alignment can produce, in addition to 1-1 links, 2-1, 1-2, 2-2, 0-1, and 1-0 links. In manual verification, a number of other links were discovered as well. First, where there was a sequence of 0-1 or 1-0 links, these were (typically) merged into 0-n or n-0 links. Such links were due to translators not translating a portion of the text. But furthermore, other link arities were discovered, e.g. 1-6 and 2-4 links. The table below summarizes all the link arities encountered in the six translation-original alignments of MULTEXT-East:

Link	BG-EN	CS-EN	ET-EN	HU-EN	RO-EN	SL-EN
0 - 1	16	21	2	19	10	3
0 - 2	0	1	1	3	2	0
0 - 3	0	0	0	0	2	0
0 - 4	0	0	0	1	0	0
1 - 0	0	2	1	1	0	2
1 - 1	6623	6439	6428	6477	6047	6572
1 - 2	36	78	100	47	259	53
1 - 3	0	2	1	0	14	0
1 - 4	0	0	0	0	0	0
1 - 5	0	0	0	1	1	1
1 - 6	0	0	0	1	0	0
2 - 1	22	110	58	108	85	48
2 - 2	2	0	3	0	2	0
2 - 3	0	0	0	0	3	0
2 - 4	0	0	0	0	1	0
3 - 1	0	2	2	7	3	0
3 - 3	0	0	0	0	0	1
4 - 1	0	1	0	1	0	0

Link arities in "1984" alignment

In the following section we give the details of the CES encoding of the alignment documents. As these documents do not contain the aligned sentences directly, a HTML version of the alignments was also prepared. To produce it, the NSL software produced by LTG was used. For details on this software see *http://www.ltg.ed.ac.uk/software/*.

2.1 Corpus Encoding

Each of the six pairwise alignments with English has been encoded as a separate SGML document, with the <**cesAlign**> root element, and stored in a separate file.

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 alignment documents are given the following PUBLIC identifiers:

```
-//MTE//DOCUMENT ALIGN 1984//BG EN
-//MTE//DOCUMENT ALIGN 1984//CS EN
-//MTE//DOCUMENT ALIGN 1984//ET EN
-//MTE//DOCUMENT ALIGN 1984//HU EN
-//MTE//DOCUMENT ALIGN 1984//RO EN
-//MTE//DOCUMENT ALIGN 1984//SL EN
```

It should be noted that the bilingual specifications of the above identifiers could be invalid, as SGML OTR 9401:1997 does not make provisions for multilingual documents.

The <**cesAlign**> documents do not contain the primary data, but only links to S-level elements, expressed as pairs of ID references to the parallel S-units of the two aligned cesDoc documents.

The structure of the MULTEXT-East alignment documents is as follows: The alignment documents currently do not have a header. The root element is <**cesAlign**>, which contains a <**linkList**> element. This element can have a number of link group <**linkGrp**> elements. In MULTEXT-East each document has only one link group of **type=body** that encompasses the complete bodies of the two aligned documents. The link group contains the actual links between the S-elements, with the references to IDs given as the value of its **xTargets** attribute. This value is composed of two sequences of IDs, separated by semicolon. The ID references in the sequence are separated by spaces.

The following hypothetical Slovene-English Orwell illustrates the overall structure of an MULTEXT-East alignment document; each link gives one type (one, many, zero) of possible alignment:

```
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD cesAlign//EN">
<cesAlign version="4.1">
  <linkList id="Oslen">
    <linkGrp id="Oslen.1" type="body" targtype="s" domains="Osl Oen">
      <link xtargets="Osl1.1 ; Oen1.1">
      <link xtargets="Osl1.2 Osl1.3 ; Oenl1.2">
      <link xtargets="Osl1.4 ; ">
    </linkGrp>
  </linkList>
</cesAlign>
```

As can be seen, the only link group in the link list is of type BODY, its target type is of type S, and its domains are the Slovene and English Orwell. The first link represents an 1 - 1 alignment, the second a 2 - 1 alignment, and the third a 1 - 0 alignment.

To simplify ID reference parsing, each ID in the first sequence is followed by exactly one space, while each ID in the second sequence is preceded by one space.

Chapter 3

Morphosyntactic Tagging

In addition to the cesDoc encoding, the Orwell corpus is also available as a tokenised and morphosyntactically tagged cesAna document. For more informations on morphosyntactic descriptions (MSDs) with which the text was annotated, see the MULTEXT-East Deliverable D1.1 F. The lexica used in the annotation process are described in the Deliverable D1.2. The cesAna DTD is explained in the CES documentation.

To arrive at the tokenised and tagged cesAna Orwell the following steps have been performed:

1. the cesDoc version has been simplified and converted to cesAna encoding;
2. the text was tokenised
3. the tokens were annotated with lexical, i.e. ambiguous MSDs, lemmas and tags
4. the lexical information was disambiguated.

To explain the structure of the final documents, first consider a fragment of the English cesDoc Orwell:

```
<!DOCTYPE text PUBLIC "-//CES//DTD cesDoc//EN">
<text>
  <body lang=en id=Oen>
    <div id="Oen.1" type=part n=1>
      <div id="Oen.1.1" type=chapter n=1>
        <p id="Oen.1.1.1">
          <s id="Oen.1.1.1.1">
            It was a bright cold day in April,
            and the clocks were striking thirteen.
          </s>
          <s id="Oen.1.1.1.2">
            <name type=person>Winston Smith</name>,
            his chin nuzzled into his breast in an effort to escape the
            vile wind, slipped quickly through the glass doors of
            <name type=place>Victory Mansions</name>,
```



```
            though not quickly enough to prevent a swirl of gritty dust
            from entering along with him.
          </s>
        </p>
...
```

The gross document structure of a cesDoc document is different from the cesAna one. In the Orwell corpus the following relations exist between the two:

- the TEXT is encoded as CHUNKLIST
- the BODY is encoded as CHUNK
- the DIV tags are omitted
- the QUOTE tags are omitted
- the P-level elements are encoded as PAR elements:
  - P is PAR, with implied TYPE;
  - the HEAD elements or omitted,
    if present they are encoded as PAR TYPE=HEAD
  - LIST and POEM elements can be omitted,
    if present they are encoded as PAR TYPE=LIST and TYPE=POEM respectively
- the S-level elements are encoded as S elements:
  - S is S, with implied TYPE;
  - if ITEM and L are present, they are marked as TYPE=ITEM and TYPE=L.
  - P-level and S-level IDs are referred to in the FROM attribute of PAR and S.
- the Q tags are omitted
- other cesDoc (sub-S level) tags such as DATE, NAME, ABBR, etc., are encoded as values of the CLASS attribute of the TOKen element.

Follows an example the derived cesAna from the above cesDoc, marked according to these conventions:

```
<!DOCTYPE cesAna PUBLIC "-//CES//DTD cesAna//EN">
<cesAna version="4.6">
  <chunkList type=TEXT>
   <chunk  type=BODY>
     <par from="Oen.1.1.1">
      <s from="Oen.1.1.1.1">
        It was a bright cold day in April,
        and the clocks were striking thirteen.
      </s>
      <s from="Oen.1.1.1.2">
```

```
        Winston Smith
        his chin nuzzled into his breast in an effort to escape the
        vile wind, slipped quickly through the glass doors of
        Victory Mansions,
        though not quickly enough to prevent a swirl of gritty dust
        from entering along with him.
        </s>
     </par>
...
```

At the S level the documents have been tokenised according the lexical resources of the language and are encoded as TOKen elements. Tokens are either 'normal' words, compounds, separable parts of words ('clitics'), or punctuation marks. They are distinguished by the value of the token's TYPE attribute. The values used are WORD for words, and PUNCT for punctuation marks. The word or punctuation mark is contained in the ORTH element. The punctuation tokens are annotated with (unambiguous) corpus tags, which identical across the languages of MULTEXT-East; for a description of the tags used, see the MULTEXT-East Deliverable D1.1 F, Section 2.5.1. The following example illustrates this markup:

```
 <chunkList type=TEXT lang=en>
  <chunk type=BODY lang=en>
   <par from='Oen.1.1.1'>
    <s from='Oen.1.1.1.1'>
     <tok type=WORD><orth>It</orth></tok>
     <tok type=WORD><orth>was</orth></tok>
     <tok type=WORD><orth>a</orth></tok>
     <tok type=WORD><orth>bright</orth></tok>
     <tok type=WORD><orth>cold</orth></tok>
     <tok type=WORD><orth>day</orth></tok>
     <tok type=WORD><orth>in</orth></tok>
     <tok type=WORD><orth>April</orth></tok>
     <tok type=PUNCT><orth>,</orth><ctag>COMMA</ctag></tok>
     <tok type=WORD><orth>and</orth></tok>
     <tok type=WORD><orth>the</orth></tok>
     <tok type=WORD><orth>clocks</orth></tok>
     <tok type=WORD><orth>were</orth></tok>
     <tok type=WORD><orth>striking</orth></tok>
     <tok type=WORD><orth>thirteen</orth></tok>
     <tok type=PUNCT><orth>.</orth><ctag>PERIOD</ctag></ctag>
     </tok>
    </s>
```

The word tokens are annotated both with ambiguous lexical information, and with context-dependent, disambiguated information. The former is contained in the <**lex**> elements of the token, the latter in the <**disamb**> element(s). Both elements contain the <**base**> (lemma) of the token, its morphosyntactic description <**msd**>, and (depending on the language) its corpus tag, <**ctag**>, as illustrated in the following example:


```
<tok type=WORD>
 <orth>glass</orth>
 <disamb><base>glass</base><msd>Afp</msd><ctag>ADJE</ctag></disamb>
 <lex><base>glass</base><msd>Afp</msd><ctag>ADJE</ctag></lex>
 <lex><base>glass</base><msd>Ncns</msd><ctag>NN</ctag></lex>
</tok>
<tok type=WORD>
 <orth>doors</orth>
 <disamb><base>door</base><msd>Ncnp</msd><ctag>NNS</ctag></disamb>
 <lex><base>door</base><msd>Ncnp</msd><ctag>NNS</ctag></lex>
</tok>
```

It should be noted that the correct annotation is given both in the <**disamb**>, as well as in (one of) the <**lex**> elements. The <**lex**> elements of token thus represent its ambiguity class. In generally, however, there may be more <**disamb**> elements for one token, in cases where the tagger or human could not decide how to disambiguate. In this case each <**disamb**> element appears among the <**lex**> elements as well.

GI	**EN**	**BG**	**CS**	**ET**	**HU**	**RO**	**SL**
par	1,286	1,322	1,297	1,266	1,303	1,343	1,288
s	6,701	6,682	6,751	6,478	6,768	6,521	6,689
tok	118,102	101,173	100,358	94,906	98,426	118,063	107,770
orth	118,102	101,173	100,358	94,906	98,426	118,063	107,770
disamb	187,526	86,020	79,862	75,433	80,705	101,508	90,792
lex	214,404	156,002	214,368	147,542	111,945	189,695	187,562
base	401,930	242,022	294,230	222,975	192,650	291,203	278,354
msd	401,930	156,002	294,230	222,975	192,650	291,203	278,354
ctag	416,035	257,175	20,496	94,906	98,426	307,758	16,978

Tag usage in Orwell's "1984"

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

The next sections give the <**cesHeader**> of the seven cesAna documents.

3.1 English

```
<cesHeader
  version="4.1"
  type="text"
  lang=en
  creator=ET
  status="update"
```

```
  date.created="1997-12-15"
  date.updated="1997-12-20"
>
    <filedesc>
      <titlestmt>
        <h.title>Multext-East cesAna: Nineteen Eighty-Four, English</h.title>
        <respstmt>
          <respname>Nancy Ide</respname>
          <resptype>Overall Responsibility</resptype>
          <respname>Greg Priest-Dorman</respname>
          <resptype>Generation of Lexical Data</resptype>
          <respname>Vladim&iacute;r Petkevi&ccaron;</respname>
          <resptype>Conversion to cesAna DTD</resptype>
        </respstmt>
      </titlestmt>
      <editionstmt version="1.0">MTE Final Release</editionstmt>
      <extent>
        <wordCount>103997</wordCount>
        <byteCount units="MB">cca 38</byteCount>
        <extnote>wordCount represents the number of TOK TYPE=WORD
          elements in the text.</extnote>
      </extent>
      <publicationstmt>
        <distributor>
          Department of Computer Science, Vassar College
        </distributor>
        <pubaddress>Poughkeepsie, New York 12604-0252  USA</pubaddress>
        <eaddress type="email">ide@cs.vassar.edu</eaddress>
        <availability status="restricted">
          Available for research purposes upon receipt of signed agreement
        </availability>
        <pubDate value="1998-01-01">January 1st, 1998</pubDate>
      </publicationstmt>
      <sourcedesc>
        <biblfull>
          <titlestmt>
            <h.title>Multext-East CES1: Nineteen Eighty-Four, English</h.title>
          </titlestmt>
          <publicationstmt>
            <distributor>
              Department of Computer Science, Vassar College
            </distributor>
            <pubaddress>Poughkeepsie, New York 12604-0252  USA</pubaddress>
            <eaddress type="email">ide@cs.vassar.edu</eaddress>
            <availability status="restricted">
              Available for research purposes upon receipt of signed agreement
            </availability>
            <pubDate value="1997-10-01">October 1, 1997</pubDate>
          </publicationstmt>
      <sourcedesc>
        <biblfull>
          <titlestmt>
            <h.title>
```



```
              The European Corpus Initiative
              Multilingual Corpus 1:
              1984 by George Orwell (English)
            </h.title>
            <respstmt>
              <respname>Association for Computational Linguistics</respname>
              <resptype>Converted from OTA's DTD to ECI DTD</resptype>
            </respstmt>
          </titlestmt>
          <publicationstmt>
            <distributor>ACL</distributor>
            <pubaddress>ACL</pubaddress>
            <availability status=restricted>
              Available for research purposes upon receipt of signed
              agreement
            </availability>
            <pubdate>1994</pubdate>
          </publicationstmt>
          <sourcedesc>
            <biblfull>
              <titlestmt>
                <h.title>Orwell's 1984: electronic edition</h.title>
                <respstmt>
                  <respname>Oxford Text Archive</respname>
                  <resptype>
                    The four versions of Orwell's 1984 in the OTA
                    were all 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 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. The other languages were
                    converted to TEI conformant SGML by the ECI
                    project 1993.) --LB, Nov 1992
                  </resptype>
                </respstmt>
              </titlestmt>
              <editionstmt>
                Public Domain TEI edition prepared at the Oxford Text
                Archive
              </editionstmt>
              <publicationstmt>
                <distributor>Oxford Text Archive</distributor>
                <pubaddress>
                  Oxford University Computing Service
                  13 Banbury Road
                  Oxford OX2 6NN UK
                  archive@ox.ac.uk
                </pubaddress>
```

```
            <availability status=restricted>
              Freely available for non-commercial
              use provided that this header is included in its
              entirety with any copy distributed
            </availability>
            <pubdate>19 Nov 1992</pubdate>
          </publicationstmt>
          <sourcedesc>
            <biblstruct>
              <monogr>
                <h.title>1984</h.title>
                <h.author>George Orwell</h.author>
                <imprint>
                  <pubdate>1949; reprinted 1961</pubdate>
                  <publisher>New American Library</publisher>
                  <pubplace>New York</pubplace>
                </imprint>
              </monogr>
            </biblstruct>
          </sourcedesc>
        </biblfull>
      </sourcedesc>
    </biblfull>
  </sourcedesc>
  </biblfull>
  </sourcedesc>
</filedesc>
<encodingdesc>
  <projectdesc>
    MULTEXT-East:
    Multilingual Text Tools and Corpora for Central and Eastern
    European Languages.
    EU Copernicus Project COP106
  </projectdesc>
  <editorialdecl>
    <transduction>
      In the cesDoc to cesAna conversion, DIV, QUOTE, Q tags and
      HEAD, POEM, LIST elements have been omitted. cesDoc P
      elements are encoded as PAR, and S as S.
      cesDoc sub-S level tags are omitted: DATE, NAME, ABBR, etc.
    </transduction>
    <quotation>
      Q and QUOTE tags from the cesDoc source not retained.
    </quotation>
    <segmentation>
      S segmentation same as in cesDoc source (hand-validated).
      TOK segmentation performed with mtseg and manually corrected,
     </segmentation>
  </editorialdecl>
  <tagsdecl>
    <tagusage gi=chunklist occurs=1>
      Element corresponds to TEXT of the cesDoc source
    </tagusage>
```



```
    <tagusage gi=chunk occurs=1>
      Element corresponds to BODY of the cesDoc source
    </tagusage>
    <tagusage gi=par occurs=1286>
      Elements correspond to P elements of the cesDoc source.
      The FROM attribute gives the reference to the ID of the
      corresponding cesDoc P element.
    </tagusage>
    <tagusage gi=s occurs=6701>
      Elements correspond to S elements of the cesDoc source
      The FROM attribute gives the reference to the ID of the
      corresponding cesDoc S element.
    </tagusage>
    <tagusage gi=tok occurs=118102>
      Tokens are of TYPE=WORD or PUNCT, with the CLASS attribute
      giving the mtseg class of the token (ABBR, COMP, INIT, TTL).
      The FROM attribute gives reference to the ID of the
      corresponding cesDoc S element in which the token in
      question appears along with the character offset of the
      token within the sentence (the character offset is appended
      to the sentence ID).
    </tagusage>
    <tagusage gi=orth occurs=118102>
      Contains the orthography of the token, as found in the cesDoc source
       (except for COMP, which have underscore instead of blank).
    </tagusage>
      <tagusage gi=disamb occurs=187526>
        Contains disambiguated lexical information for
        WORDs. Disambiguation performed by Eric Brill's Unsupervised
        Part-of-Speech Tagger Version 0.8.  Trained on chapters 1&2
        of Multext-East CES1: Nineteen Eighty-Four, English.  A token
        with several DISAMBs indicates that the Brill Tagger was
        not able to fully disambiguate the token.  In such a case
        all equally-weighted possibilities are listed.
      </tagusage>
      <tagusage gi=lex occurs=214404>
        Contains undisambiguated lexical information for WORDs.
      </tagusage>
      <tagusage gi=base occurs=401930>
        Base or lemmma of a WORD. In the event that the base of
        the WORD was not known, the content of this tag will be
        "??" (two question marks).
      </tagusage>
      <tagusage gi=msd occurs=401930>
        Morphosyntactic description of a WORD. In the event that
        the MSD of the WORD was not known, the content of this tag
        will be "??" (two question marks).
      </tagusage>
      <tagusage gi=ctag occurs=416035>
        Corpus tag (for tok type=WORD and for tok type=PUNCT).
        In the event that the CTAG of the WORD was not known, the
        content of this tag will be "??" (two question marks).
      </tagusage>
```

```
    </tagsdecl>
  </encodingdesc>
  <profiledesc>
    <creation date="1997-12-15"></creation>
    <langusage>
      <![ %ONECOMPONENT [ &ISOlang; ]]>
      <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>
    </langusage>
  </profiledesc>
  <revisiondesc>
    <change>
      <changedate>1997-12-16</changedate>
      <respname>Vladim&iacute;r Petkevi&ccaron;</respname>
      <h.item>Revised several tagusage descriptions, and supplied
              counts in the header.
      </h.item>
    </change>
     <change>
     <changedate>1997-12-20</changedate>
      <respname>Tomaz Erjavec, IJS</respname>
       <h.item>Modified EDITIONSTMT</h.item>
     </change>
  </revisiondesc>
</cesHeader>
```

3.2 Bulgarian

```
<cesHeader
  version="4.1"
  type="text"
  lang=en
  creator=LD
  status="update"
  date.created="1997-11-30"
  date.updated="1997-12-21"
>
  <filedesc>
   <titlestmt>
     <h.title>Multext-East cesAna: Nineteen Eighty-Four, Bulgarian</h.title>
       <respstmt>
         <respname>Ludmila Dimitrova, Lydia Sinapova</respname>
         <resptype>Overall Responsibility</resptype>
         <respname>Ludmila Dimitrova, Kiril Simov</respname>
         <resptype>Hand-tagging of first chapter first part</resptype>
         <respname>Ludmila Dimitrova</respname>
         <resptype>Hand-tagging of second chapter first part, first
           chapter second part</resptype>
         <respname>Vladim&iacute;r Petkevi&ccaron;</respname>
         <resptype>Conversion to cesAna DTD </resptype>
```



```
       </respstmt>
       </titlestmt>
       <editionstmt version="1.0">MTE Final Release</editionstmt>
       <extent>
         <wordCount>86020</wordCount>
         <byteCount units="MB">29.9</byteCount>
         <extnote>wordCount represents the number of TOK TYPE=WORD
            elements in the text.
         </extnote>
       </extent>
       <publicationstmt>
         <distributor>
           Institute of Mathematics and Informatics,
           Bulgarian Academy of Sciences, Sofia
         </distributor>
         <pubaddress>
           Acad G. Bonchev st. bl.8
           1113 Sofia, Bulgaria
         </pubaddress>
         <eaddress type="email">ludmila@ling.math.acad.bg</eaddress>
         <availability status="restricted">
           Available for research purposes upon receipt of signed
           agreement
         </availability>
         <pubDate value="1998-01-01">January 1st, 1998</pubDate>
       </publicationstmt>
       <sourcedesc>
         <biblfull>
           <titlestmt>
             <h.title>Multext-East CES1: Nineteen Eighty-Four, Bulgarian
             </h.title>
           </titlestmt>
           <publicationstmt>
             <distributor>
              Institute of Mathematics and Informatics,
              Bulgarian Academy of Sciences, Sofia
             </distributor>
             <pubaddress>
              Acad G. Bonchev st. bl.8
              1113 Sofia, Bulgaria
             </pubaddress>
             <eaddress type="email">ludmila@ling.math.acad.bg</eaddress>
             <availability status="restricted">
               Available for research purposes upon receipt of signed
               agreement
             </availability>
             <pubDate value="1997-10-01">October 1, 1997</pubDate>
           </publicationstmt>
       <sourcedesc>
         <biblfull>
           <titlestmt>
             <h.title> Electronic form of 1984 by George Orwell in
                       Bulgarian
```

```
        </h.title>
            <respstmt>
              <respname>
                Ludmila Dimitrova (BAS), Lydia Sinapova (BAS),
                Kiril Simov(BAS)
              </respname>
              <resptype>
                 Typing-in 1984.
              </resptype>
            </respstmt>
      </titlestmt>
      <publicationstmt>
        <distributor>
         Institute of Mathematics and Informatics,
         Bulgarian Academy of Sciences, Sofia
        </distributor>
        <pubaddress>
         Acad G. Bonchev st. bl.8
         1113 Sofia, Bulgaria
        </pubaddress>
        <availability status=restricted>
          Available for research purposes upon receipt of signed
          agreement
        </availability>
        <pubdate>1997</pubdate>
      </publicationstmt>
          <sourcedesc>
            <biblstruct>
              <monogr>
                <h.title>1984)</h.title>
                <h.author>George Orwell</h.author>
                <h.author>Translator: Lydia Bozhilova</h.author>
                <imprint>
                  <pubdate>1989</pubdate>
                  <publisher>Profizdat</publisher>
                  <pubplace>Sofia, Bulgaria</pubplace>
                </imprint>
              </monogr>
            </biblstruct>
          </sourcedesc>
        </biblfull>
      </sourcedesc>
    </biblfull>
  </sourcedesc>
 </filedesc>
<encodingdesc>
  <projectdesc>
    MULTEXT-East:
    Multilingual Text Tools and Corpora for Central and Eastern
    European Languages.
    EU Copernicus Project COP106
  </projectdesc>
  <editorialdecl>
```



```
    <transduction>
      In the cesDoc to cesAna conversion, DIV, QUOTE, Q tags and
      HEAD, POEM, LIST elements have been omitted. cesDoc P
      elements are encoded as PAR, and S as S.
      cesDoc sub-S level tags are omitted: DATE, NAME, ABBR, etc.
    </transduction>
    <quotation>
      Q and QUOTE tags from the cesDoc source not retained.
    </quotation>
    <segmentation>
      S segmentation same as in cesDoc source (hand-validated).
      TOK segmentation performed with mtseg and manually corrected,
     </segmentation>
  </editorialdecl>
  <tagsdecl>
    <tagusage gi=chunklist occurs=1>
      Element corresponds to TEXT of the cesDoc source
    </tagusage>
    <tagusage gi=chunk occurs=1>
      Element corresponds to BODY of the cesDoc source
    </tagusage>
   <tagusage gi=par occurs=1322>
      Elements correspond to P elements of the cesDoc source.
      The FROM attribute gives the reference to the ID of the
      corresponding cesDoc P element.
    </tagusage>
    <tagusage gi=s occurs=6682>
      Elements correspond to S elements of the cesDoc source
      The FROM attribute gives the reference to the ID of the
    corresponding cesDoc S element.
    </tagusage>
    <tagusage gi=tok occurs=101173>
      Tokens are of TYPE=WORD or PUNCT, with the CLASS attribute
      giving the mtseg class of the token.
    </tagusage>
    <tagusage gi=orth   occurs=101173>
      Contains the orthography of the token, as found in the
      cesDoc source.
    </tagusage>
    <tagusage gi=disamb occurs=86020>
      Contains disambiguated lexical information.
    </tagusage>
    <tagusage gi=lex    occurs=156002>
      Contains undisambiguated lexical information.
    </tagusage>
    <tagusage gi=base   occurs=242022>
      Base or lemma of a token.
    </tagusage>
   <tagusage gi=msd    occurs=156002>
      Morphosyntactic description of a token.
    </tagusage>
    <tagusage gi=ctag   occurs=257175>
      Corpus tag.
```

```
      </tagusage>
    </tagsdecl>
  </encodingdesc>
  <profiledesc>
    <creation date="1997-11-27">
    </creation>
    <langusage>
      <![ %ONECOMPONENT [ &ISOlang; ]]>
      <language id=ns-bg iso639=bg>Newspeak Bulgarian</language>
    </langusage>
  </profiledesc>
  <revisiondesc>
     <change>
     <changedate>1997-12-19</changedate>
      <respname>Vladim&iacute;r Petkevi&ccaron;, &Uacute;TKL FFUK,
                Prague</respname>
       <h.item>Filled in tags' usage, wordcount and bytecount</h.item>
     </change>
     <change>
     <changedate>1997-12-21</changedate>
      <respname>Tomaz Erjavec, IJS</respname>
       <h.item>Converted from ISO Cyrillic to SGML entities</h.item>
       <h.item>Changed ... to &hellip;</h.item>
       <h.item>Modified EDITIONSTMT, BYTECOUNT</h.item>
     </change>
   </revisiondesc>
 </cesheader>
```


3.3 Czech


```
<cesHeader
 version="4.1"
 type="text"
 lang=en
 creator=VP
 status="update"
 date.created="1997-11-28"
 date.updated="1997-12-21"
>
   <filedesc>
     <titlestmt>
       <h.title>Multext-East cesAna: Nineteen Eighty-Four, Czech</h.title>
       <respstmt>
         <respname>Vladim&iacute;r Petkevi&ccaron;</respname>
         <resptype>Overall Responsibility</resptype>
         <respname>Milena Hn&aacute;tkov&aacute;</respname>
         <resptype>Hand-tagging of the first 3 chapters</resptype>
         <resptype>Revision of the tagger results</resptype>
         <respname>Vladim&iacute;r Petkevi&ccaron;</respname>
         <resptype>Conversion to cesAna DTD</resptype>
       </respstmt>
```



```
     </titlestmt>
     <editionstmt version="1.0">MTE Final Release</editionstmt>
     <extent>
       <wordCount>79862</wordCount>
       <byteCount units="MB">24.4</byteCount>
     <extnote>
        wordCount represents he number of TOK TYPE=WORD
        elements in the text.
     </extnote>
     </extent>
     <publicationstmt>
       <distributor>
         Institute of Theoretical and Computational Linguistics,
         Faculty of Philosophy, Charles University, Prague
       </distributor>
       <pubaddress>Celetn&aacute; 13 110 00 Praha 1, Czech Republic</pubaddress>
       <eaddress type="email">Vladimir.Petkevic@ff.cuni.cz</eaddress>
       <availability status="restricted">
         Available for research purposes upon receipt of signed agreement
       </availability>
       <pubDate value="1998-01-01">January 1st, 1998</pubDate>
     </publicationstmt>
     <sourcedesc>
       <biblfull>
         <titlestmt>
           <h.title>Multext-East CES1: Nineteen Eighty-Four, Czech</h.title>
         </titlestmt>
         <publicationstmt>
           <distributor>
             Institute of Theoretical and Computational Linguistics,
             Faculty of Philosophy, Charles University, Prague
           </distributor>
         <pubaddress>Celetn&aacute; 13, 110 00 Praha 1, Czech Republic</pubaddress>
         <eaddress type="email">Vladimir.Petkevic@ff.cuni.cz</eaddress>
         <availability status="restricted">
           Available for research purposes upon receipt of signed agreement
         </availability>
         <pubDate value="1997-10-01">November 1, 1997</pubDate>
       </publicationstmt>
     <sourcedesc>
       <biblfull>
         <titlestmt>
           <h.title>
             Electronic form of 1984 by George Orwell in Czech,
             obtained via OCR
           </h.title>
           <respstmt>
             <respname>
               Vladim&iacute;r Petkevi&ccaron;
               Institute of Theoretical and Computational Linguistics,
               Faculty of Philosophy, Charles University, Prague, Czech Republic
               (&Uacute;TKL FFUK)
             </respname>
```

```
        <resptype>
          OCR'ed the novel
        </resptype>
      </respstmt>
    </titlestmt>
    <publicationstmt>
      <distributor>
        Institute of Theoretical and Computational Linguistics,
        Faculty  of Philosophy, Charles University, Prague, Czech Republic
        (&Uacute;TKL FFUK)
      </distributor>
      <pubaddress>
        Celetn&aacute; 13, Praha 1
        Czech Republic
      </pubaddress>
      <availability status=restricted>
        Available for research purposes upon receipt of signed
        agreement
      </availability>
      <pubdate>1998</pubdate>
    </publicationstmt>
  <sourcedesc>
    <biblstruct>
      <monogr>
        <h.title>1984</h.title>
        <h.author>George Orwell</h.author>
        <h.author>Translator: Eva &Scaron;ime&ccaron;kov&aacute;</h.author>
        <imprint>
          <pubdate>1991</pubdate>
          <publisher>Na&scaron;e vojsko</publisher>
          <pubplace>Prague, Czech Republic</pubplace>
        </imprint>
      </monogr>
    </biblstruct>
  </sourcedesc>
    </biblfull>
  </sourcedesc>
    </biblfull>
  </sourcedesc>
</filedesc>
<encodingdesc>
  <projectdesc>
    MULTEXT-East:
    Multilingual Text Tools and Corpora for Central and Eastern
    European Languages.
    EU Copernicus Project COP106
  </projectdesc>
  <editorialdecl>
    <transduction>
      In the cesDoc to cesAna conversion, DIV, QUOTE, Q tags and
      HEAD, POEM, LIST elements have been omitted. cesDoc P
      elements are encoded as PAR, and S as S.
      cesDoc sub-S level tags are omitted: DATE, NAME, ABBR, etc.
```



```
    </transduction>
    <quotation>
       Q and QUOTE tags from the cesDoc source not retained.
    </quotation>
    <segmentation>
       S segmentation same as in cesDoc source (hand-validated).
       TOK segmentation performed with mtseg and manually corrected,
    </segmentation>
  </editorialdecl>
  <tagsdecl>
    <tagusage gi=chunklist occurs=1>
      Element corresponds to TEXT of the cesDoc source
    </tagusage>
    <tagusage gi=chunk occurs=1>
      Element corresponds to BODY of the cesDoc source
    </tagusage>
    <tagusage gi=par occurs=1297>
      Elements correspond to P elements of the cesDoc source.
      The FROM attribute gives the reference to the ID of the
      corresponding cesDoc P element.
    </tagusage>
    <tagusage gi=s occurs=6751>
      Elements correspond to S elements of the cesDoc source
      The FROM attribute gives the reference to the ID of the
      corresponding cesDoc S element.
    </tagusage>
    <tagusage gi=tok occurs=100358>
      Tokens are of TYPE=WORD or PUNCT, with the CLASS attribute
      giving the mtseg class of the token.
    </tagusage>
    <tagusage gi=orth   occurs=100358>
      Contains the orthography of the token, as found in the
      cesDoc source.
    </tagusage>
    <tagusage gi=disamb occurs=79862>
      Contains disambiguated lexical information.
    </tagusage>
    <tagusage gi=lex    occurs=214368>
      Contains undisambiguated lexical information.
    </tagusage>
    <tagusage gi=base   occurs=294230>
      Base or lemma of a token.
    </tagusage>
    <tagusage gi=msd    occurs=294230>
      Morphosyntactic description of a token.
    </tagusage>
    <tagusage gi=ctag   occurs=20496>
      Corpus tag (only for punctuation).
    </tagusage>
  </tagsdecl>
</encodingdesc>
<profiledesc>
  <creation date="1997-11-04">
```

```
      </creation>
      <langusage>
        <![ %ONECOMPONENT [ &ISOlang; ]]>
      <language id=ns-cs iso639=cs>Newspeak Czech</language>
      </langusage>
    </profiledesc>
    <revisiondesc>
      <change>
        <changedate>1997-11-04</changedate>
        <respname>Toma&zcaron; Erjavec, IJS</respname>
        <h.item>Created initial header template and part of the content</h.item>
      </change>
      <change>
      <changedate>1997-11-28</changedate>
        <respname>Vladim&iacute;r Petkevi&ccaron;, &Uacute;TKL</respname>
        <h.item>Created the specific part of the header content</h.item>
      </change>
      <change>
      <changedate>1997-12-21</changedate>
       <respname>Tomaz Erjavec, IJS</respname>
        <h.item>Converted from ISO Latin-2 to SGML entities</h.item>
        <h.item>Modified EDITIONSTMT, BYTECOUNT</h.item>
      </change>
    </revisiondesc>
</cesheader>
```

3.4 Estonian

```
<cesHeader
  version="4.1"
  type="text"
  lang=en
  creator=HJK
  status="update"
  date.created="1997-11-28"
  date.updated="1997-12-21"
>
    <filedesc>
      <titlestmt>
        <h.title>Multext-East cesAna: Nineteen Eighty-Four, Estonian</h.title>
        <respstmt>
          <respname>Heiki-Jaan Kaalep</respname>
          <resptype>Overall Responsibility</resptype>
          <respname>Kadri Muischnek</respname>
          <resptype>Hand-tagging of part 1, chapter 1-4;
                    part 2 chapter 9</resptype>
          <respname>Andriela R&auml;&auml;bis</respname>
          <resptype>Hand-tagging of part 1, chapter 5-7;
                    part 3 chapter 1, 3, 4</resptype>
          <respname>Heili Orav</respname>
          <resptype>Hand-tagging of part 1, chapter 8;
```



```
                    part 3 chapter 2, 5, 6</resptype>
          <respname>Helen Potter</respname>
          <resptype>Hand-tagging of part 2, chapter 1-7</resptype>
          <respname>K&uuml;lli Habicht</respname>
          <resptype>Hand-tagging of part 2, chapter 8</resptype>
         <respname>Vladim&iacute;r Petkevi&ccaron;</respname>
          <resptype>Conversion to cesAna DTD </resptype>
        </respstmt>
      </titlestmt>
      <editionstmt version="1.0">MTE Final Release</editionstmt>
      <extent>
        <wordCount>75433</wordCount>
        <byteCount units="MB">18.7 MB</byteCount>
        <extnote>wordCount represents he number of TOK TYPE=WORD
          elements in the text. byteCount is in megaBytes</extnote>
      </extent>
      <publicationstmt>
        <distributor>
           T&Uuml; arvutuslingvistika uurimisgrupp
        </distributor>
        <pubaddress>Tiigi 78-232, Tartu, Estonia</pubaddress>
        <eaddress type="email">hkaalep@psych.ut.ee</eaddress>
        <eaddress type="www">http://www.cl.ut.ee</eaddress>
        <availability status="free">
           Freely available
        </availability>
        <pubDate value="1998-01-01">January 1st, 1998</pubDate>
      </publicationstmt>
      <sourcedesc>
        <biblfull>
          <titlestmt>
            <h.title>Multext-East CES1: Nineteen Eighty-Four, Estonian</h.title>
          </titlestmt>
          <publicationstmt>
            <distributor>
           T&Uuml; arvutuslingvistika uurimisgrupp
            </distributor>
            <pubaddress>Tiigi 78-232, Tartu, Estonia</pubaddress>
            <eaddress type="email">hkaalep@psych.ut.ee</eaddress>
            <eaddress type="www">http://www.cl.ut.ee</eaddress>
            <availability status="free">
             Freely available
            </availability>
            <pubDate value="1997-10-01">October 1, 1997</pubDate>
          </publicationstmt>
              <sourcedesc>
                <biblstruct>
                  <monogr>
                    <h.title>1984</h.title>
                    <h.author>George Orwell</h.author>
                    <h.author>Translator: Elias Treeman</h.author>
                    <imprint>
                      <pubdate>1990</pubdate>
```

```
                <publisher>Loomingu Raamatukogu nr. 48-51</publisher>
                <publisher>Perioodika</publisher>
                <pubplace>Tallinn</pubplace>
              </imprint>
            </monogr>
          </biblstruct>
        </sourcedesc>
      </biblfull>
    </sourcedesc>
 </filedesc>
<encodingdesc>
  <projectdesc>
    MULTEXT-East:
    Multilingual Text Tools and Corpora for Central and Eastern
    European Languages.
    EU Copernicus Project COP106
  </projectdesc>
  <editorialdecl>
    <transduction>
      In the cesDoc to cesAna conversion, DIV, QUOTE tags and
      HEAD, POEM, LIST elements have been omitted. cesDoc P
      elements are encoded as PAR, and S as S.
      Q tags have been encoded as punctuation symbols.
      cesDoc sub-S level tags are omitted: DATE, NAME, ABBR, etc.
    </transduction>
    <quotation>
      QUOTE tags from the cesDoc source not retained.
    </quotation>
    <segmentation>
      S segmentation same as in cesDoc source (hand-validated).
      TOK segmentation performed with mtseg and manually corrected,
     </segmentation>
  </editorialdecl>
  <tagsdecl>
    <tagusage gi=chunkList occurs=1>
      Element corresponds to TEXT of the cesDoc source
    </tagusage>
    <tagusage gi=chunk occurs=1>
      Element corresponds to BODY of the cesDoc source
    </tagusage>
    <tagusage gi=par occurs=1266>
      Elements correspond to P elements of the cesDoc source.
      The FROM attribute gives the reference to the ID of the
      corresponding cesDoc P element.
    </tagusage>
    <tagusage gi=s occurs=6478>
      Elements correspond to S elements of the cesDoc source
      The FROM attribute gives the reference to the ID of the
    corresponding cesDoc S element.
    </tagusage>
    <tagusage gi=tok occurs=94906>
      Tokens are of TYPE=WORD or PUNCT, with the CLASS attribute
      giving the mtseg class of the token.
```



```
    </tagusage>
    <tagusage gi=orth   occurs=94906>
      Contains the orthography of the token, as found in the
      cesDoc source.
    </tagusage>
    <tagusage gi=disamb occurs=75433>
      Contains disambiguated lexical information.
    </tagusage>
    <tagusage gi=lex    occurs=147542>
      Contains undisambiguated lexical information.
    </tagusage>
    <tagusage gi=base   occurs=222975>
      Base or lemmma of a token.
    </tagusage>
    <tagusage gi=msd    occurs=222975>
      Morphosyntactic description of a token.
    </tagusage>
    <tagusage gi=ctag   occurs=94906>
      Corpus tag.
    </tagusage>
  </tagsdecl>
</encodingdesc>
<profiledesc>
  <creation date="1997-11-28">
  </creation>
  <langusage>
    <![ %ONECOMPONENT [ &ISOlang; ]]>
  </langusage>
</profiledesc>
<revisiondesc>
  <change>
    <changedate>1997-11-28</changedate>
    <respname>Heiki-Jaan Kaalep</respname>
    <h.item>Initial header</h.item>
  </change>
   <change>
   <changedate>1997-12-21</changedate>
    <respname>Tomaz Erjavec, IJS</respname>
     <h.item>Modified EDITIONSTMT and changed ... to &hellip;</h.item>
   </change>
</revisiondesc>
</cesheader>
```

3.5 Hungarian

```
<cesHeader
  version="4.1"
  type="text"
  lang=en
  creator=OCS
  status="update"
```

```
  date.created="1997-11-24"
  date.updated="1997-12-21"
>
    <filedesc>
      <titlestmt>
        <h.title>Multext-East cesAna: Nineteen Eighty-Four, Hungarian</h.title>
        <respstmt>
          <respname>Csaba Oravecz</respname>
          <resptype>Overall Responsibility</resptype>
          <respname>Vladim&iacute;r Petkevi&ccaron;</respname>
          <resptype>Conversion to cesAna DTD </resptype>
        </respstmt>
      </titlestmt>
      <editionstmt version="1.0">MTE Final Release</editionstmt>
      <extent>
        <wordCount>80705</wordCount>
        <byteCount units="MB">18.4</byteCount>
        <extnote>wordCount represents he number of TOK TYPE=WORD
          elements in the text. byteCount is in megaBytes</extnote>
      </extent>
      <publicationstmt>
        <distributor>
           Research Institute for Linguistics, Hungarian Academy of Sciences
        </distributor>
        <pubaddress> Budapest, Sz&iacute;nh&aacute;z u. 5-9.</pubaddress>
        <eaddress type="email">oravecz@nytud.hu</eaddress>
        <eaddress type="www">http://www.nytud.hu</eaddress>
        <availability status="restricted">
          Available for research purposes upon receipt of signed agreement
        </availability>
        <pubDate value="1998-01-01">January 1st, 1998</pubDate>
      </publicationstmt>
      <sourcedesc>
         <biblStruct>
         <monogr>
           <h.title>1984</h.title>
           <h.author>George Orwell</h.author>
           <imprint>
           <pubdate>1989</pubdate>
           <publisher>Eur&oacute;pa K&ouml;nyvkiad&oacute;</publisher>
           <pubplace>Budapest</pubplace>
           </imprint>
         </monogr>
         </biblStruct>
      </sourcedesc>
    </filedesc>
    <encodingdesc>
      <projectdesc>
        MULTEXT-East:
        Multilingual Text Tools and Corpora for Central and Eastern
        European Languages.
        EU Copernicus Project COP106
      </projectdesc>
```



```
      <editorialdecl>
        <transduction>
          In the cesDoc to cesAna conversion, DIV, QUOTE, Q tags and
          HEAD, POEM, LIST elements have been omitted. cesDoc P
          elements are encoded as PAR, and S as S.
          cesDoc sub-S level tags are omitted: DATE, NAME, ABBR, etc.
        </transduction>
        <quotation>
          Q and QUOTE tags from the cesDoc source not retained.
        </quotation>
        <segmentation>
          S segmentation same as in cesDoc source (hand-validated).
          TOK segmentation performed with mtseg and manually corrected,
        </segmentation>
      </editorialdecl>
      <tagsdecl>
        <tagusage gi=chunkList occurs=1>
          Element corresponds to TEXT of the cesDoc source
        </tagusage>
        <tagusage gi=chunk occurs=1>
          Element corresponds to BODY of the cesDoc source
        </tagusage>
        <tagusage gi=par occurs=1303>
          Elements correspond to P elements of the cesDoc source.
          The FROM attribute gives the reference to the ID of the
          corresponding cesDoc P element.
        </tagusage>
        <tagusage gi=s occurs=6768>
          Elements correspond to S elements of the cesDoc source
          The FROM attribute gives the reference to the ID of the
        corresponding cesDoc S element.
        </tagusage>
        <tagusage gi=tok occurs=98426>
          Tokens are of TYPE=WORD or PUNCT, with the CLASS attribute
          giving the mtseg class of the token.
        </tagusage>
        <tagusage gi=orth   occurs=98426>
          Contains the orthography of the token, as found in the
          cesDoc source.
        </tagusage>
        <tagusage gi=disamb occurs=80705>
          Contains disambiguated lexical information.
        </tagusage>
        <tagusage gi=lex    occurs=111945>
          Contains undisambiguated lexical information.
        </tagusage>
        <tagusage gi=base   occurs=192650>
          Base or lemmma of a token.
        </tagusage>
        <tagusage gi=msd    occurs=192650>
          Morphosyntactic description of a token.
        </tagusage>
        <tagusage gi=ctag   occurs=98426>
```

```
        Corpus tag.
      </tagusage>
    </tagsdecl>
  </encodingdesc>
  <profiledesc>
    <creation date="1997-11-04">
    </creation>
    <langusage>
      <![ %ONECOMPONENT [ &ISOlang; ]]>
      <language id=ns-hu iso639=hu>Newspeak Hungarian</language>
    </langusage>
  </profiledesc>
  <revisiondesc>
    <change>
      <changedate>1997-11-24</changedate>
      <respname>Csaba Oravecz, RIL</respname>
      <h.item>Initial header</h.item>
    </change>
     <change>
     <changedate>1997-12-21</changedate>
      <respname>Tomaz Erjavec, IJS</respname>
       <h.item>Converted from ISO Latin-2 to SGML entities</h.item>
       <h.item>Changed ... to &hellip;</h.item>
       <h.item>Modified EDITIONSTMT, BYTECOUNT</h.item>
     </change>
  </revisiondesc>
</cesheader>
```

3.6 Romanian

```
<cesHeader
  version="4.1"
  type="text"
  lang=en
  creator=DT
  status="update"
  date.created="1997-11-04"
  date.updated="1997-12-21"
>
    <filedesc>
      <titlestmt>
        <h.title>Multext-East cesAna: Nineteen Eighty-Four, Romanian</h.title>
        <respstmt>
          <respname>Dan Tufi&scedil;</respname>
          <resptype>Overall Responsibility</resptype>
          <respname>Ana-Maria Barbu</respname>
          <resptype>Hand-tagging the whole book</resptype>
          <respname>Vasile P&abreve;tra&scedil;cu </respname>
          <resptype>Conversion to cesAna DTD </resptype>
        </respstmt>
      </titlestmt>
```



```
      <editionstmt version="1.0">MTE Final Release</editionstmt>
       <extent>
        <wordCount>101508</wordCount>
        <byteCount units="MB">27.1</byteCount>
        <extnote>wordCount represents he number of TOK TYPE=WORD
          elements in the text. byteCount is in megaBytes</extnote>
      </extent>
       <publicationstmt>
        <distributor>
          Center for Advanced Research in Machine Learning, Natural  Language
                  Processing and Conceptual Modelling
        </distributor>
        <pubaddress>Casa Academiei,13, "13 Septembrie, Bucharest 5, 74311, Romania"
                <pubaddress>
        <eaddress type="email">tufis@valhalla.racai.ro</eaddress>
        <eaddress type="www">http://nl.ijs.si/ME</eaddress>
        <availability status="restricted">
          Available for research purposes upon receipt of signed agreement
        </availability>
        <pubDate value="1998-01-01">January 1st, 1998</pubDate>
       </publicationstmt>
       <sourcedesc>
         <biblstruct>
          <monogr>
           <h.title>O mie nou&abreve; sute optzeci &scedil;i patru</h.title>
           <h.author>George Orwell</h.author>
           <h.author>Translator: Mihnea Gafi&tcedil;a</h.author>
           <imprint>
            <pubdate>1991</pubdate>
            <publisher>Editura Univers</publisher>
            <pubplace>Bucharest</pubplace>
           </imprint>
          </monogr>
         </biblstruct>
       </sourcedesc>
    </filedesc>
    <encodingdesc>
     <projectdesc>
        MULTEXT-East:
        Multilingual Text Tools and Corpora for Central and Eastern
        European Languages.
        EU Copernicus Project COP106
     </projectdesc>
     <editorialdecl>
        <transduction>
          The electronic form was obtained by keyboarding at the
          Center for Advanced Research in Machine Learning, Natural  Language
          Processing and Conceptual Modelling, spell-checked and hand tagged.
          In the cesDoc to cesAna conversion, DIV, QUOTE, Q tags and
          HEAD, POEM, LIST elements have been omitted. cesDoc P
          elements are encoded as PAR, and S as S.
          cesDoc sub-S level tags are omitted: DATE, NAME, ABBR, etc.
        </transduction>
```

```
    <quotation>
      Q and QUOTE tags from the cesDoc source not retained.
    </quotation>
   <segmentation>
      S segmentation same as in cesDoc source (hand-validated).
      TOK segmentation performed with mtseg and manually corrected,
     </segmentation>
  </editorialdecl>
  <tagsdecl>
    <tagusage gi=chunkList occurs=1>
      Element corresponds to TEXT of the cesDoc source
    </tagusage>
    <tagusage gi=chunk occurs=1>
      Element corresponds to BODY of the cesDoc source
    </tagusage>
    <tagusage gi=par occurs=1343>
      Elements correspond to P, POEM, LIST, HEAD elements of the cesDoc source.
      The FROM attribute gives the reference to the ID of the
      corresponding cesDoc P element.
    </tagusage>
    <tagusage gi=s occurs=6521>
      Elements correspond to S, L, ITEM elements of the cesDoc source
      The FROM attribute gives the reference to the ID of the
    corresponding cesDoc S element.
    </tagusage>
    <tagusage gi=tok occurs=118063>
      Tokens are of TYPE=WORD or PUNCT, with the CLASS attribute
      giving the mtseg class of the token.
    </tagusage>
    <tagusage gi=orth   occurs=118063>
      Contains the orthography of the token, as found in the
      cesDoc source.
    </tagusage>
    <tagusage gi=disamb occurs=101508>
      Contains disambiguated lexical information.
    </tagusage>
    <tagusage gi=lex    occurs=189695>
      Contains undisambiguated lexical information.
    </tagusage>
    <tagusage gi=base   occurs=291203>
      Base or lemmma of a token.
    </tagusage>
    <tagusage gi=msd    occurs=291203>
      Morphosyntactic description of a token.
    </tagusage>
    <tagusage gi=ctag   occurs=307758></tagusage>
  </tagsdecl>
   </encodingdesc>
<profiledesc>
  <creation date="1997-11-04">
  </creation>
  <langusage>
    <![ %ONECOMPONENT [ &ISOlang; ]]>
```



```
      <language id=ns-ro iso639=sl>Newspeak Romanian</language>
    </langusage>
  </profiledesc>
  <revisiondesc>
    <change>
      <changedate>1997-11-04</changedate>
      <respname>Toma&zcaron; Erjavec, IJS</respname>
      <h.item>Initial header</h.item>
    </change>
    <change>
      <changedate>1997-11-06</changedate>
      <respname>Vasile P&abreve;tra&scedil;cu</respname>
       <h.item>
        The Tagusage, Bytecount and Wordcount were updated. Entities that
        were counted as words are those that were identified by the segmenter
        that is words, clitics, compounds (counted as one unit, irrespective
        of the number of constituents), punctuation, numbers.
      </h.item>
    </change>
     <change>
     <changedate>1997-12-21</changedate>
      <respname>Tomaz Erjavec, IJS</respname>
       <h.item>Modified EDITIONSTMT and changed ... to &hellip;</h.item>
     </change>
  </revisiondesc>
</cesheader>
```

3.7 Slovene

```
<cesHeader
  version="4.1"
  type="text"
  lang=en
  creator=ET
  status="update"
  date.created="1997-11-04"
  date.updated="1997-12-20"
>
    <filedesc>
      <titlestmt>
        <h.title>Multext-East cesAna: Nineteen Eighty-Four, Slovene</h.title>
        <respstmt>
          <respname>Toma&zcaron; Erjavec</respname>
          <resptype>Overall Responsibility</resptype>
          <respname>Aleksandra Bizjak, Primo&zcaron; Jakopin</respname>
          <resptype>Tagging</resptype>
          <respname>Vladim&iacute;r Petkevi&ccaron;</respname>
          <resptype>Conversion to cesAna DTD </resptype>
        </respstmt>
      </titlestmt>
      <editionstmt version="1.0">MTE Final Release</editionstmt>
```

```
<extent>
  <wordCount>90768</wordCount>
  <byteCount units="MB">22.5</byteCount>
  <extnote>wordCount represents the number of TOK TYPE=WORD
    elements in the text.</extnote>
</extent>
<publicationstmt>
  <distributor>
    Dept. for Intelligent Systems, Jozef &Scaron;tefan Institute
  </distributor>
  <pubaddress>Jamova 39, SI-1000 Ljubljana, Slovenia</pubaddress>
  <eaddress type="email">tomaz.erjavec@ijs.si</eaddress>
  <eaddress type="www">http://nl.ijs.si/ME</eaddress>
  <availability status="restricted">
    Available for research purposes upon receipt of signed agreement
  </availability>
  <pubDate value="1998-01-01">January 1st, 1998</pubDate>
</publicationstmt>
<sourcedesc>
  <biblfull>
    <titlestmt>
      <h.title>Multext-East CES1: Nineteen Eighty-Four, Slovene</h.title>
    </titlestmt>
    <publicationstmt>
      <distributor>
        Dept. for Intelligent Systems, Jozef &Scaron;tefan Institute
      </distributor>
      <pubaddress>Jamova 39, SI-1000 Ljubljana, Slovenia</pubaddress>
      <eaddress type="email">tomaz.erjavec@ijs.si</eaddress>
      <eaddress type="www">http://nl.ijs.si/ME</eaddress>
      <availability status="restricted">
        Available for research purposes upon receipt of signed agreement
      </availability>
      <pubDate value="1997-10-01">October 1, 1997</pubDate>
    </publicationstmt>
<sourcedesc>
  <biblfull>
    <titlestmt>
      <h.title>
        The European Corpus Initiative
        Multilingual Corpus 1:
        1984 by George Orwell (Slovene)
      </h.title>
      <respstmt>
        <respname>Association for Computational Linguistics</respname>
        <resptype>Converted from OTA's DTD to ECI DTD</resptype>
      </respstmt>
    </titlestmt>
    <publicationstmt>
      <distributor>ACL</distributor>
      <pubaddress>ACL</pubaddress>
      <availability status=restricted>
        Available for research purposes upon receipt of signed
```



```
        agreement
      </availability>
      <pubdate>1994</pubdate>
    </publicationstmt>
    <sourcedesc>
      <biblfull>
        <titlestmt>
          <h.title>Orwell's 1984: electronic edition</h.title>
          <respstmt>
            <respname>Oxford Text Archive</respname>
            <resptype>
              The four versions of Orwell's 1984 in the OTA
              were all 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 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. The other languages were
              converted to TEI conformant SGML by the ECI
              project 1993.) --LB, Nov 1992
            </resptype>
          </respstmt>
        </titlestmt>
        <editionstmt>
          Public Domain TEI edition prepared at the Oxford Text
          Archive
        </editionstmt>
        <publicationstmt>
          <distributor>Oxford Text Archive</distributor>
          <pubaddress>
            Oxford University Computing Service
            13 Banbury Road
            Oxford OX2 6NN UK
            archive@ox.ac.uk
          </pubaddress>
          <availability status=restricted>
            Freely available for non-commercial
            use provided that this header is included in its
            entirety with any copy distributed
          </availability>
          <pubdate>19 Nov 1992</pubdate>
        </publicationstmt>
        <sourcedesc>
          <biblstruct>
            <monogr>
              <h.title>1984</h.title>
              <h.author>George Orwell</h.author>
              <h.author>Translator: Alenka Puhar</h.author>
              <imprint>
```

```
         <pubdate>1983</pubdate>
         <publisher>Knji&zcaron;nica Kondor</publisher>
         <publisher>Mladinska knjiga</publisher>
         <pubplace>Ljubljana</pubplace>
        </imprint>
       </monogr>
      </biblstruct>
     </sourcedesc>
    </biblfull>
   </sourcedesc>
  </biblfull>
 </sourcedesc>
 </biblfull>
 </sourcedesc>
</filedesc>
<encodingdesc>
 <projectdesc>
  MULTEXT-East:
  Multilingual Text Tools and Corpora for Central and Eastern
  European Languages.
  EU Copernicus Project COP106
 </projectdesc>
 <editorialdecl>
  <transduction>
   In the cesDoc to cesAna conversion, DIV, QUOTE, Q tags and
   HEAD, POEM, LIST elements have been omitted. cesDoc P
   elements are encoded as PAR, and S as S.
   cesDoc sub-S level tags are omitted: DATE, NAME, ABBR, etc.
  </transduction>
  <quotation>
   Q and QUOTE tags from the cesDoc source not retained.
  </quotation>
  <segmentation>
   S segmentation same as in cesDoc source (hand-validated).
   TOK segmentation performed with mtseg and manually corrected,
  </segmentation>
 </editorialdecl>
 <tagsdecl>
  <tagusage gi=chunklist occurs=1>
   Element corresponds to TEXT of the cesDoc source
  </tagusage>
  <tagusage gi=chunk occurs=1>
   Element corresponds to BODY of the cesDoc source
  </tagusage>
  <tagusage gi=par occurs=1288>
   Elements correspond to P elements of the cesDoc source.
   The FROM attribute gives the reference to the ID of the
   corresponding cesDoc P element.
  </tagusage>
  <tagusage gi=s occurs=6689>
   Elements correspond to S elements of the cesDoc source
   The FROM attribute gives the reference to the ID of the
   corresponding cesDoc S element.
```



```
   </tagusage>
   <tagusage gi=tok occurs=107770>
    Tokens are of TYPE=WORD or PUNCT, with the CLASS attribute
    giving the mtseg class of the token (ABBR, COMP, INIT, TTL).
   </tagusage>
   <tagusage gi=orth occurs=107770>
    Contains the orthography of the token, as found in the cesDoc source
     (except for COMP, which have underscore instead of blank).
   </tagusage>
   <tagusage gi=disamb occurs=90792>
    Contains disambiguated lexical information for WORDs.
   </tagusage>
   <tagusage gi=lex occurs=187562>
    Contains undisambiguated lexical information for WORDs.
   </tagusage>
   <tagusage gi=base occurs=278354>
    Base or lemmma of a WORD.
   </tagusage>
   <tagusage gi=msd occurs=278354>
    Morphosyntactic description of a WORD.
   </tagusage>
   <tagusage gi=ctag occurs=16978>
    Corpus tag of PUNCT tokens.
   </tagusage>
  </tagsdecl>
 </encodingdesc>
 <profiledesc>
  <creation date="1997-12-11"></creation>
  <langusage>
   <![ %ONECOMPONENT [ &ISOlang; ]]>
   <language id=ns-sl iso639=sl>Newspeak Slovene</language>
  </langusage>
 </profiledesc>
 <revisiondesc>
  <change>
   <changedate>1997-11-04</changedate>
   <respname>Toma&zcaron; Erjavec, IJS</respname>
   <h.item>Initial header</h.item>
  </change>
  <change>
   <changedate>1997-12-11</changedate>
   <respname>Toma&zcaron; Erjavec, IJS</respname>
   <h.item>Manually disambiguated version prepared</h.item>
  </change>
  <change>
   <changedate>1997-12-20</changedate>
   <respname>Toma&zcaron; Erjavec, IJS</respname>
   <h.item>Prepared final version</h.item>
  </change>
 </revisiondesc>
</cesHeader>
```