Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages

Ludmila DIMITROVA
Institute of Mathematics and Informatics
Sofia, Bulgaria
ludmila@ling.math.acad.bg

Tomaz ERJAVEC
Institute Jozef Stefan
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Nancy IDE
Dept, of Computer Science, Vassar College
Poughkeepsie, New York, USA
ide@cs.vassar.edu

Heiki Jaan KAALEP
Dept. of General Linguistics , University of Tartu
Tartu, Estonia
hkaalep@psych.ut.ee

Vladimir PETKEVIC
Inst. of Theoretical and Computational
Linguistics, Charles University
Prague, Czech Republic
Vladimir.Petkevic@ff.cuni.cz

Dan TUFIS
Romanian Academy
Center for Artificial Intelligence
Bucharest, Romania
tufis@racai.ro

Abstract

The EU Copernicus project Multext-East has created a multi-lingual corpus of text and speech data, covering the six languages of the project: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene. In addition, wordform lexicons for each of the languages were developed. The corpus includes a parallel component consisting of Orwell's *Nineteen Eighty-Four*, with versions in all six languages tagged for part-of-speech and aligned to English (also tagged for POS). We describe the encoding format and data architecture designed especially for this corpus, which is generally usable for encoding linguistic corpora. We also describe the methodology for the development of a harmonized set of morphosyntactic descriptions (MSDs), which builds upon the scheme for western European languages developed within the EAGLES project. We discuss the special concerns for handling the six project languages, which cover three distinct language families.

Introduction

In order to provide resources to enable the efficient extraction of quantitative and qualitative information from corpora, several corpus development and distribution efforts have been recently established. However, few corpora exist for Central and Eastern European (CEE) languages, and corpus-processing tools that take into account the specific characteristics of these languages are virtually non-existent.

The Multext-East Copernicus project[1] (Erjavec, et al., 1997) was a spin-off of the LRE project Multext[2] (Ide and Véronis, 1994) intended to fill these gaps by developing significant resources for six CEE languages (Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene) that follow a consistent and principled encoding format and are maximally suited to easy processing by corpus-handling tools. To this end, Multext-East developed a corpus of parallel and comparable texts for the six CEE project languages, together with wordform lexicons and other language-specific resources. In the following sections we briefly describe the Multext-East corpora (text, speech) and the Multext-East lexicons and language-specific resources.

1 The Multext-East corpora

1.1 Encoding format

Based on the principle that its corpus encoding format should be standardized and homogeneous both for interchange and for facilitating open-ended retrieval tasks, Multext-East adopted the

[1] http://nl.ijs.si/ME

[2] http://www.lpl.univ-aix.fr/projects/Multext/

Corpus Encoding Standard (CES)[3] (Ide, 1998), which has been developed to be optimally suited for use in language engineering and corpus-based work. The CES is an application of SGML (ISO-8879, Standard Generalized Markup Language) and is based on the *TEI Guidelines for Electronic Text Encoding and Interchange.*
In addition to providing encoding conventions for elements relevant to corpus-based work, the CES provides a data architecture for linguistic corpora and their annotations. Each corpus component, comprising a single text and its annotations, is organized as a *hyper-document*, with various levels of annotation stored in separate SGML documents (each with a separate DTD). Low-density (i.e., above the token level) annotation is expressed indirectly in terms of inter-document links. Markup for different types of annotation (e.g., part of speech, alignment, etc.) is described by a separate Data Type Definition (DTD) specifically tailored to that information.

1.2 The parallel corpus

The Multext-East parallel corpus consists of seven translations of George Orwell's *Nineteen Eighty-Four*: besides the original English version, the corpus contains translations in the six project languages. There are three versions of each text in the parallel corpus, corresponding to different levels of annotation: a cesDoc encoding (SGML markup up to the sub-paragraph level, including markup for sentence boundaries); and a cesAna encoding, containing word-level morphosyntactic markup together with links to each sentence (and in some versions, to each word) in the cesDoc version. A fourth document, the cesAlign document, is associated with each of the non-English versions, which includes links between sentences in the cesDoc encoding for each and the English version, thus providing a parallel alignment at the sentence level. The cesAna versions, which are the most linguistically informative, are marked up as shown below for the English phrase "smell of bugs":

```
<tok type=WORD from='Oen.1.6.15.1\62'>
  <orth>smell</orth>
  <disamb><base>smell</base><msd>Ncns
    </msd><ctag>NN</ctag></disamb>
  <lex><base>smell</base>
       <msd>Vmip-p</msd>
       <ctag>VERB</ctag></lex>
  <lex><base>smell</base>
       <msd>Vmip1s</msd>
       <ctag>VERB</ctag></lex>
  <lex><base>smell</base>
       <msd>Vmip2s</msd>
       <ctag>VERB</ctag></lex>
  <lex><base>smell</base>
       <msd>Vmn</msd>
       <ctag>VINF</ctag></lex>
  <lex><base>smell</base>
       <msd>Ncns</msd>
       <ctag>NN</ctag></lex></tok>
<tok type=WORD from='Oen.1.6.15.1\68'>
  <orth>of</orth>
  <disamb><base>of</base>
          <msd>Sp</msd>
          <ctag>PREP</ctag></disamb>
  <lex><base>of</base>
       <msd>Sp</msd>
       <ctag>PREP</ctag></lex></tok>
<tok type=WORD from='Oen.1.6.15.1\71'>
  <orth>bugs</orth>
  <disamb><base>bug</base>
          <msd>Ncnp</msd>
          <ctag>NNS</ctag></disamb>
  <lex><base>bug</base>
       <msd>Vmip3s</msd>
       <ctag>VERB3</ctag></lex>
  <lex><base>bug</base>
       <msd>Ncnp</msd>
       <ctag>NNS</ctag></lex></tok>
```

In this example, the position of each token in the parallel corpus is given in the *from* attribute whose value specifies the hierarchical position of the token within the text (here, the token "smell" appears in part 1, chapter 6, paragraph 15, sentence 1, byte offset 62). All possible morphosyntactic interpretations of the token are given in the `<lex>` field consisting of the base form, a morphosyntactic description (see Section 2), and an associated corpus tag. The `<disamb>` field contains the interpretation that has been identified as valid within the respective context; within this tag, the `<ctag>` element provides the corresponding *corpus tag* (see section 2).[4] The

[3] The CES was developed in a joint effort of the European projects Multext (LRE) and EAGLES (in particular, the EAGLES Text Representation subgroup), together with the Vassar/CNRS collaboration (supported by the U.S. National Science Foundation).

[4] In the Czech and Slovene versions, <ctag> is omitted because its contents are identical to the <msd> tag contents.

disambiguation of each language version in the parallel corpus was accomplished using automatic POS tagging algorithms and then partially or entirely hand-validated.

Table 1 provides the main characteristics per language of this corpus. In this table:

- *tok* = number of tokens
- *words* = number of lexical items (excluding punctuation)
- *lex* = number of MSD-based interpretations of the words in the text
- *MSD/amb* = average ratio of the number of lexical variants per word

The texts from the corpora were segmented using the corpus annotation toolset developed within the Multext project, augmented by language-specific resources developed by Multext-East. The Multext segmenter is a language-independent and configurable tokenizer whose output includes token, paragraph and sentence boundary markers. Punctuation, lexical items, numbers, and various alphanumeric sequences (such as dates and hours) are annotated with tags defined in a hierarchical, class-structured tagset. The language-specific behavior of the segmenter is enabled by its engine-driven design, in which all language-specific information is provided as data. Within Multext-East, resource data, including rules describing the form of sentence boundaries, word splitting (cliticized forms decomposition), word compounding, quotations, numbers, dates, punctuation, capitalization, abbreviations etc., was developed for the six project languages.

Once the input text was tokenized, a dictionary look-up procedure was used to assign each lexical token all its possible morphosyntactic descriptors (MSDs). The ambiguously MSD-annotated texts were then hand-disambiguated (entirely for some languages and partially for the others). This time-consuming and error-prone process was sped up significantly by a special XEMACS mode, developed within the project, which is aware of the morphosyntactic descriptors' significance and allows for natural language expansion of the linear encoding of the MSDs. The ambiguously MSD-annotated texts and the corresponding disambiguated texts provided the basis for building the cesAna encoded version of the multilingual parallel corpus.

The corpus also contains six language pair-wise alignments between each of the six project languages and English. The alignments were performed by three different automatic aligners (Multext-aligner, "vanilla-aligner", Silfide-aligner) with accuracy ranging between 75-90%, and then hand validated. Table 2 shows the distribution of sentence alignments for each pair of languages.

1.3 Multilingual comparable corpus

Multext-East also produced a multilingual *comparable* corpus, including two subsets of at least 100,000 words each for each of the six project languages. The texts include fiction, comprising a single novel or excerpts from several novels, and newspaper data.

The data is comparable across the six languages, in terms of the number and size of texts. The entire multilingual comparable corpus was prepared in CES format manually or using *ad hoc* tools.

Language	Bulgarian	Czech	English	Estonian	Hungarian	Romanian	Slovene
tokens	101173	100358	118102	94906	98426	118063	107769
words	86020	79862	103997	75433	80705	101508	90792
lex	156002	214368	214404	147542	111945	189695	187562
MSD/amb	1.81	2.68	2.06	1.96	1.39	1.87	2.07
distinct words	16348	19115	9745	17870	20316	15225	17861
distinct lemmas	8517	9161	7260	8873	10387	7433	7916

Table 1: Corpus characteristics

Finno-Ugric Languages | **Romance Language**

Estonian-English

Align	Nr.	Proc
3-1	2	0.030321%
2-2	3	0.045482%
2-1	60	0.909642%
1-3	1	0.015161%
1-2	100	1.516070%
1-1	6426	97.422680%
1-0	1	0.015161%
0-2	1	0.015161%
0-1	2	0.030321%

Hungarian-English

Align	Nr.	Proc
7-0	1	0.014997%
4-1	1	0.014997%
3-1	7	0.104979%
3-0	1	0.014997%
2-1	108	1.619676%
1-6	1	0.014997%
1-5	1	0.014997%
1-2	46	0.689862%
1-1	6479	97.165567%
0-4	1	0.014997%
0-2	3	0.044991%
0-1	19	0.284943%

Romanian-English

Align	Nr.	Proc
3-1	3	0.046656%
2-4	1	0.015552%
2-3	3	0.046656%
2-2	2	0.031104%
2-1	85	1.321928%
2-0	1	0.015552%
1-5	1	0.015552%
1-3	14	0.217729%
1-2	259	4.027994%
1-1	6047	94.043546%
0-3	2	0.031104%
0-2	2	0.031104%
0-1	10	0.155521%

Slavic Languages

Bulgarian-English

Align	Nr.	Proc
2-2	2	0.030017%
2-1	23	0.345190%
1-2	72	1.080594%
1-1	6558	98.424133%
0-1	8	0.120066%

Czech-English

Align	Nr.	Proc
4-1	1	0.015029%
3-1	2	0.030057%
2-1	109	1.638112%
1-3	2	0.030057%
1-2	81	1.217313%
1-1	6438	96.753832%
0-1	21	0.315600%

Slovene-English

Align	Nr.	Proc
3-3	1	0.014970%
2-1	48	0.718563%
1-5	1	0.014970%
1-2	53	0.793413%
1-1	6572	98.383234%
1-0	2	0.029940%
0-1	3	0.044910%

Table 2: Distribution of sentence alignments

2 Morpho-lexical resources

Multext-East, in collaboration with EAGLES, evaluated, adapted and extended the EAGLES morphosyntactic specifications (rule format, lexical specifications, corpus tagset, etc.) to cover the six Multext-East languages (Erjavec and Monachini, 1997). Accommodating the different language families represented among the Multext-East languages demanded substantial assessment and modification of the pre-existing specifications, which were originally developed for western European languages only.

For corpus morpho-lexical processing purposes, the Multext-East consortium developed language-specific wordform dictionaries, which, for all languages except Estonian and Hungarian, contain the full inflectional paradigm for at least the lemmas appearing in the corpus. Each dictionary entry has the following structure:

```
wordform [TAB] lemma [TAB] MSD [TAB]
```

where *wordform* represents an inflected form of the lemma, characterised by a combination of feature values encoded by a Morphosyntactic Description (MSD). The Multext-East lexicons and MSDs are fully described in Tufis, Ide, and Erjavec (1998).

A general overview of the lexicons is shown in Table 3. The **Entries** column provides the number of dictionary entries, that is, triplets: <wordform lemma MSD>. The **Wordforms** column gives the number of distinct wordforms appearing in the lexicon, irrespective of their lemma and MSD. The **Lemma** column gives the number of distinct lemmas in the lexicon, eliminating duplications that appear due to lemma homography. The difference between the **Lemma** and "**=**" fields provides an estimate of the number of homographic lemmas. The MSD field gives the total number of distinct MSDs used in the encoding of the lexicon stock.

The last two columns in Table 3 (AMB_POS and AMB_MSD) provide information about the number of *ambiguity classification clusters*. An ambiguity classification cluster provides the number of ways a homographic wordform can be classified. AMB_POS ("part of speech ambiguity") and AMB_MSD ("MSD-ambiguity") provide the classification based on the part of speech and MSD, respectively. The number of ambiguity classes (based either on POS or MSD) is a key figure in estimating the space needed to construct a statistical language

model (such as HMM) useful for morphosyntactic disambiguation. This number was a key factor in the tagset design.

For several of the project languages and for English, a set of *corpus tags* has also been developed which are appropriate for use with stochastic disambiguators. Where corpus tags have been developed, mapping rules from MSDs to corpus tags (n-to-1 mapping) are also provided as a resource.

Language	Entries	Wordforms	Lemmas	=	MSD	AmbPOS	AmbMSD
English	66469	43455	22571	25813	132	47	248
Romanian	440363	347960	33259	35421	674	90	981
Slovene	539213	191728	15671	15863	2044	48	1185
Czech	133803	41601	14458	14684	915	35	698
Bulgarian	333779	284211	18864	19071	185	42	400
Estonian	130409	89180	22054	23384	563	63	1012
Hungarian	59614	46886	15838	17380	603	62	890

Table 3: Multilingual Lexicon Overview

Conclusion

The multilingual resources (lexicons, rules, corpora) developed in Multext-East are among of the most comprehensive resources currently available for most of the project languages. In addition to resource development, the work carried out in Multext-East has contributed significantly to defining general mechanisms for lexical specification, and it has provided a test of the extensibility of standards and tools beyond the languages for which they were originally developed. All Multext-East resources and tools are distributed, at cost, on CD ROM through the TELRI project[5] (see Erjavec, Lawson, and Romary, 1998).

Acknowledgements

This project was supported by EU Copernicus Project COP106. The U.S. portion was supported in part by NSF grant IRI-9413451. We would like to thank the following for their contribution to the project: G. Priest-Dorman, A.M. Barbu, C. Popescu, V. Patrascu, G. Rotariu, S. Bruda, J. Véronis, S. Harié, P. DiCristo, L. Sinapova, R. Pavlov, K. Simov, D. Popov, S. Vidinska, M. Hnatkova, J. Hajic, B. Hladka, A. Bizjak, P .Jakopin, M. Romih, O. Vukovic, and M. Boldea. We would also like to acknowledge Laboratoire Parole et Langage, CNRS, Aix-en-Provence, which coordinated the project.

References

Erjavec, T., Ide, N., and Tufis, D. (1998) *Standardized Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages.* First International Language Resources and Evaluation Conference, Granada, Spain.

Erjavec, T. and Monachini, M. (Eds.) (1997). *Specifications and Notation for Lexicon Encoding.* Deliverable D1.1 F. Multext-East Project COP-106. http://nl.ijs.si/ME/CD/docs/mte-d11f/.

Erjavec, T., Ide, N., Petkevic, P. and Véronis, J. (1996). *Multext-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages*. Proceedings of the Trans European Language Resource Infrastructure First Conference, Tihany, pp. 87--98.

Ide, N. (1998) *Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora* First International Language Resources and Evaluation Conference, Granada, Spain. See also http://www.cs.vassar.edu/CES/.

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

Erjavec, T., Lawson, A. and Romary, L. (1998). *East meets West: Producing Multilingual Resources in a European Context.* First International Language Resources and Evaluation Conference, Granada, Spain.

[5] http://www.ids-mannheim.de/telri/