An annotated bibliography of the MULTEXT-East project

Tomaž Erjavec

Department of Knowledge Technologies
Jožef Stefan Institute
Jamova 39,
SI-1000 Ljubljana
Slovenia
tomaz.erjavec at ijs.si

2004-05-10

This paper overviews publications connected to the language resources produced in the MULTEXT-East project or its follow-ups:

There are so far four overview papers on MULTEXT-East: the project was presented at the outset in [21], and at its completion in [7]. The second vesion of the resources the "Concede Edition" was presented in [14]. The current, third version, bringing together the first two is published in [16], and a copy of the paper is available at mte-lrec2004.pdf.

In the context of overview papers we should also mention [33], as the most widely cited paper about MULTEXT.

Various aspects of the project at completion were presented at the LREC'98 conference: [20] dealt with the MULTEXT-East corpus, and [32] with the lexicon, while [24] introduced the double CD-ROM published by TELRI, one volume of which contained the extended MULTEXT-East deliverables. At the same conference, the Corpus Encoding Standard, CES, the encoding of the MULTEXT-East corpora, was also presented [28].

The details of the project are explained in three MULTEXT-East deliverable reports, all available on the WWW. The most substantive one is report D1.1F "Specifications and Notation for Lexicon Encoding", [40]. This report has been substantially revised and expanded in the subsequent editions of the MULTEXT-East deliverables. In particular, version 2 was presented in [14], and version 3 in [22].

The other two reports of MULTEXT-East were MTE:D21F Corpus Collection and Preparation [8] and the MTE:D23F Corpus Markup [9]. Two reports of the MULTEXT project also substantially influenced the work in MULTEXT-East, namely the report on the MtSeg tool [5] and the MULTEXT specifications for lexicon encoding, [2].

The project's morphosyntactically annotated '1984' novel represented the first tagged and available corpus for most languages involved in the project; at the same time, language independent taggers were becoming available. It is therefore not surprising that the most use was made of the Orwell corpus as a dataset for experiments in tagging models. An experiment that uses most of the multilingual corpus is [26]. The other research concentrates on particular MULTEXT-East languages, and has mostly been presented at the LREC'00 conference. Tagging Romanian has been studied in [41] and [42], Hungarian in [44] and Slovene in [12], the latter based on the more substantial report [11].

A similar strand of research into learning tagging or morphological modules has also used the MULTEXT-East corpus and lexicon. But here the methods are less statistical and fall more into the Machine Learning paradigm. A series of WWW sites has been set up, dubbed "Learning Language in Logic" that also contain bibliographies related to learning from MULTEXT-East data. Inductive Learning of Multilingual Morphology was presented in [10] and [37]. Learning to tag Slovene is discussed in [6] and learning word segmentation rules for tag prediction in [34]. Tagging Hungarian was also studied in the context of LLL and using MULTEXT-East data; details are given in [1, 27].

One of the aims of the EU Concede project was to integrate the corpus results of MULTEXT-East with lexical databases. The initial dictionary headwords were sampled from the corpus [25], and a summary of preliminary results, along with a integration of the Concede English-Slovene sample with the MULTEXT-East corpus was presented in [17]. The final results were presented in [18].

The parallel corpus was used in experiments in automatic bi-lingual lexicon extraction; the work on Romanian-English was presented in [43], in (D. Tufis et al. ALC-ACH 2001), and in Journal of Science and Technology of Information, in print.

The '1984' corpus was also used on another strand of research in which discusses cross-lingual sense determination and was published in [29], [30], [31].

Last, but not least, are the papers that discuss a particular language in the scope of MULTEXT-East or utilising MULTEXT-East resources. The Czech '1984' was discussed in [38]. A paper connected to exploitation of the Romanian portion of the resources, in the context of making a Web-based corpus server is [4]. The Slovene MULTEXT-East lexicon was presented in [13], and work on a platform using MULTEXT derived specifications for Slovene in [39]. At the same conference, work on Latvian was also reported [36]. The Slovene morphosyntactic specifications were subsequnelty used to annotate the FIDA reference corpus of the Slovene language [19].

For most of the languages in question, the original MULTEXT-East annotation work was a pioneering effort, so it was hardly surprising that during use a number of errors and inconsistencies were discovered in the data and specifications. These errors were subsequently corrected, but because the work was done at different sites and in different manners, the corpus encodings had begun to drift apart. The EU project Concede (Consortium for Central European Dictionary Encoding), [17] which ran from 1998 to 2000 and comprised most of the same partners as MULTEXT-East, offered the possibility to bring the versions back on a common footing. In the scope of a workpackage of the project, the corrected "1984" corpus was normalised and the primary data re-encoded according to the TEI (Text Encoding Initiative) guidelines and, largely, to the XML recommendation. This second version is briefly described in [15] and more fully in [14].

Finallly, Version 3 of the MULTEXT-East resources brought together both previous versions (TELRI nad CONCEDE), and made them available in XML, in TEI P4. A sampler corpus was first made from the CONCEDE release, which served as one of the test cases for the TEI Task Force on SGML to XML migration [3]. Then the complete CONCEDE and TELRI coprus was converted from SGML to XML, and from TEI P3 and CES respectivelly, to a uniform TEI P4, including the documentation. Version 3 also added components for Serbian and Resian, and is described in [16], and a copy of the paper available at mte-lrec2004.pdf.

References

[1] Zoltán Alexin, Szilvia Zvada, , and Tibor Gyimóthy. Application of AGLEARN on Hungarian Part-of-speech Tagging. In D. Parigot and M. Mernik, editors, *Second Workshop on Attribute Grammars and their Applications, WAGA'99*, pages 133–152, Amsterdam, The Netherlands, March 1999. INRIA rocquencourt. http://www-rocq.inria.fr/oscar/www/fnc2/WAGA99/accept.html.

[2] Nuria Bel, Nicoletta Calzolari, and Monica Monachini (eds.). Common specifications and notation for lexicon encoding and preliminary proposal for the tagsets. MULTEXT Deliverable D1.6.1B, ILC, Pisa, 1995.

[3] Alejandro Bia, Lou Burnard, Tomaž Erjavec, Christine Ruotolo, and Susan Schreibman. Migrating Language Resources from SGML to XML: the Text Encoding Initiative Recommendations. In *Fourth International Conference on Language Resources and Evaluation, LREC'04*, page In print, Paris, 2002. ELRA.

[4] Marian Boldea Bohus. A web-based text corpora development system. In *Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, 2000. ELRA. http://www.cs.cmu.edu/People/dbohus/docs/wbtcds_lrec2000.ps.gz.

[5] Philippe Di Cristo. Mtseg: The multext multilingual segmenter tools. MULTEXT Deliverable MSG 1, Version 1.3.1, CNRS, Aix-en-Provence, 1996. http://www.lpl.univ-aix.fr/projects/multext/MtSeg/.

[6] James Cussens, Sašo Džeroski, and Tomaž Erjavec. Morphosyntactic tagging of Slovene using Progol. In Sašo Džeroski and Peter Flach, editors, *Inductive Logic Programming: Proc. of the 9th International Workshop (ILP-99)*, number 1634 in Lecture Notes in Artificial Intelligence, pages 68–79, Bled, Slovenia, June 1999. Springer-Verlag.

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

[8] Ludmila Dimitrova, Lydia Sinapova, Vladimir Petkevič, Jana Klímová, Vera Schmiedtová, Heiki-Jan Kaalep, Viire Villandi, Heili Orav, Leho Paldre, Urve Talvik, Kadri Muischnek, Csaba Oravecz, Laszlo Tihanyi, Ştefan Bruda, Călin Diaconu, Lidia Diaconu, Dan Tufiş, Tomaž Erjavec, Miro Romih, and Olga Vukovič. Sample corpus collection and preparation. MULTEXT-East Final Report D2.1F, Institute Jožef Stefan, Ljubljana, Slovenia, December 1997. 70pp.

[9] Greg-Priest Dorman, Tomaž Erjavec, Nancy Ide, and Vladimir Petkevič. Corpus markup. MULTEXT-East Final Report D2.3F, Institute Jožef Stefan, Ljubljana, Slovenia, December 1997. 34pp.

[10] Sašo Džeroski and Tomaž Erjavec. Inductive learning of multilingual morphology. *Electrotechnical Review*, 65(6):296–302, 1998.

[11] Sašo Džeroski, Tomaž Erjavec, and Jakub Zavrel. Morphosyntactic tagging of slovene: Evaluating pos taggers and tagsets. Research Report IJS-DP 8018, Jožef Stefan Institute, Ljubljana, Slovenia, 1999. http://nl.ijs.si/lll/bib/dzerza-report/.

[12] Sašo Džeroski, Tomaž Erjavec, and Jakub Zavrel. Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. In *Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, 2000. ELRA.

[13] Tomaž Erjavec. The Multext-East Slovene Lexicon. In *Proceedings of the 7th Slovene Electrotechnical Conference, ERK '98*, pages 189–192, Portorož, Slovenia, 1998. http://nl.ijs.si/et/Bib/ERK98/.

[14] Tomaž Erjavec. Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984. In *6th Natural Language Processing Pacific Rim Symposium, NLPRS'01*, pages 487–492, Tokyo, 2001. http://nl.ijs.si/ME/V2/.

[15] Tomaž Erjavec. The MULTEXT-East Resources Revisited. *ElsNews*, 10(1):3–2, 2001.

[16] Tomaž Erjavec. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC'04*, page In print, Paris, 2004. ELRA.

[17] Tomaž Erjavec, Roger Evans, Nancy Ide, and Adam Kilgarriff. The concede model for lexical databases. In *Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, 2000. ELRA.

[18] Tomaž Erjavec, Roger Evans, Nancy Ide, and Adam Kilgarriff. From Machine Readable Dictionaries to Lexical Databases: the Concede Experience. In *Proceedings of the 7th International Conference on Computational Lexicography, COMPLEX'03*, Budapest, Hungary, 2003.

[19] Tomaž Erjavec, Vojko Gorjanc, and Marko Stabej. Korpus FIDA. In *Proceedings of the Conference 'Language Technologies for the Slovene Language'*, pages 124–127, Ljubljana, Slovenia, 1998. Institute "Jožef Stefan".

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

[21] Tomaž Erjavec, Nancy Ide, Vladimír Petkevič, and Jean Véronis. MULTEXT-East: Multilingual text tools and corpora for Central and Eastern European languages. In *Proceedings of the First TELRI European Seminar: Language Resources for Language Technology*, pages 87–98, 1996. 15–16 September 1995, Tihany, Hungary.

[22] Tomaž Erjavec, Cvetana Krstev, Vladimír Petkevič, Kiril Simov, Marko Tadić, and Duško Vitas. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*, pages 25–32, Budapest, 2003. http://nl.ijs.si/ME/V2/.

[23] Tomaž Erjavec, Ann Lawson, and Laurent Romary. East meets West: A Compendium of Multilingual Resources. CD-ROM, 1998. ISBN: 3-922641-46-6.

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

[25] Tomaž Erjavec, Dan Tufiş, and Tamas Varadi. Developing TEI-conformant lexical databases for CEE languages. In *Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX'99*, pages 205–209, Pecs, Hungary, 1999.

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

[27] T. Horváth, Zoltán Alexin, Tibor Gyimóthy, and S. Wrobel. Application of Different Learning Methods to Hungarian Part-of-speech Tagging. In Sašo Džeroski and Peter Flach, editors, *Proc. 9th Int. Conference on Inductive Logic Programming, ILP99*, number 1634 in Lecture Notes in Artificial Intelligence, pages 128–139, Bled, Slovenia, June 1999.

[28] Nancy Ide. Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 463–470, Granada, 1998. ELRA. http://www.cs.vassar.edu/CES/.

[29] Nancy Ide. Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 34(1-2):223–34, 2000.

[30] Nancy Ide, Tomaž Erjavec, and Dan Tufiş. Automatic Sense Tagging Using Parallel Corpora. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pages 83–89, Tokyo, 2001.

[31] Nancy Ide, Tomaž Erjavec, and Dan Tufiş. Sense Discrimination with Parallel Corpora. In *Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia, July 2002. ACL.

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

[33] Nancy Ide and Jean Véronis. Multext (multilingual tools and corpora). In *Proceedings of the 15th CoLing*, pages 90–96, Kyoto, 1994.

[34] Dimitar Kazakov, Suresh Manandhar, and Tomaž Erjavec. Learning word segmentation rules for tag prediction. In Sašo Džeroski and Peter Flach, editors, *Inductive Logic Programming: Proc. of the 9th International Workshop (ILP-99)*, number 1634 in Lecture Notes in Artificial Intelligence, pages 152–161, Bled, Slovenia, June 1999. Springer-Verlag.

[35] Geoffrey Leech and Andrew Wilson. Recommendations for the morphosyntactic annotation of corpora. EAGLES Report EAG–TCWG–MAC/R, ILC, Pisa, 1996. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html.

[36] Kristine Lev?ne and Andrejs Spektors. Morphemic analysis and morphological tagging of latvian corpus. In *Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, 2000. ELRA.

[37] Suresh Manandhar, Sašo Džeroski, and Tomaž Erjavec. Learning multilingual morphology with CLOG. In David Page, editor, *Inductive Logic Programming; 8th International Workshop ILP-98, Proceedings*, number 1446 in Lecture Notes in Artificial Intelligence, pages 135–144. Springer-Verlag, 1998.

[38] Vladimir Petkevič. Czech translation of G. Orwell's '1984': Morphology and syntactic patterns in the corpus. Number 1692 in Lecture Notes in Artificial Intelligence, pages 77–82. Springer-Verlag, 1999.

[39] Matej Rojc and Zdravko Kačič. A computational platform for development of morphologic and phonetic lexica. In *Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, 2000. ELRA.

[40] Specifications and notation for lexicon encoding. MULTEXT-East Final Report D1.1F, Institute Jožef Stefan, Ljubljana, Slovenia, December 1997. http://nl.ijs.si/ME/CD/docs/mte-d11f/.

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

[42] Dan Tufiş. Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as Tags for Probabilistic Tagging. In *Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, 2000. ELRA.

[43] Dan Tufiş and Ana Maria Barbu. Accurate Automatic extraction of Translation Equivalents from Parallel Corpora. In *Proceedings of the Corpus Linguistics 2001 conference*, volume 13 of *UCREL technical paper*, pages 581–586, Lancaster, 2001.

[44] Dan Tufiş, Peter Dienes, Csaba Oravecz, and Tamas Varadi. Principled hidden tagset design for tiered tagging of hungarian. In *Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, 2000. ELRA.