Lang Resources & Evaluation (2012) 46:131–142
DOI 10.1007/s10579-011-9174-8

BRIEF REPORT

MULTEXT-East: morphosyntactic resources for Central and Eastern European languages

Tomaž Erjavec

Published online: 9 December 2011

Abstract The paper presents the MULTEXT-East language resources, a multilingual dataset for language engineering research, focused on the morphosyntactic level of linguistic description. The MULTEXT-East dataset includes the morphosyntactic specifications, morphosyntactic lexica, and a parallel corpus, the novel "1984" by George Orwell, which is sentence aligned and contains hand-validated morphosyntactic descriptions and lemmas. The resources are uniformly encoded in XML, using the Text Encoding Initiative Guidelines, TEI P5, and cover 16 languages, mainly from Central and Eastern Europe: Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Resian, Romanian, Russian, Serbian, Slovak, Slovene, and Ukrainian. This dataset, unique in terms of languages covered and the wealth of encoding, is extensively documented, and freely available for research purposes. The paper overviews the MULTEXT-East resources by type and language and gives some conclusions and directions for further work.

Keywords Morphosyntactic annotation · Multilinguality · Language encoding standards

1 Introduction

The MULTEXT-East project, (Multilingual Text Tools and Corpora for Central and Eastern European Languages) ran from '95 to '97 and developed standardised language resources for six Central and Eastern European languages, as well as for English, the "hub" language of the project (Dimitrova et al. 1998). The project was

T. Erjavec (✉)
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: tomaz.erjavec@ijs.si

a spin-off of the MULTEXT project (Ide and Véronis 1994), which pursued similar goals for six Western European languages. The main results of the project were morphosyntactic specifications, defining the tagsets for lexical and corpus annotations in a common format, lexical resources and annotated multilingual corpora. In addition to delivering resources, a focus of the project was also the adoption and promotion of encoding standardization. On the one hand, the morphosyntactic annotations and lexica were developed in the formalism used in MULTEXT, itself based on the specifications of the Expert Advisory Group on Language Engineering Standards (EAGLES 1996).[1] On the other hand, the corpus resources were encoded in SGML, using CES, the Corpus Encoding Standard (Ide 1998), an application of the Text Encoding Initiative Guidelines, version P3 (Sperberg-McQueen and Burnard 1994).

After the completion of the EU MULTEXT-East project a number of further projects have helped to keep the MULTEXT-East resources up to date regarding encoding and enabled the addition of new languages. The latest release of the resources is Version 4 (V4) (Erjavec 2010), which covers 16 languages. The main improvements to Version 3 (Erjavec 2004) are the addition of resources for five new languages, updating of four, and the recoding of the morphosyntactic specifications from LaTeX to XML: the specifications, corpora, and accompanying documentation are now uniformly encoded to a schema based on the latest version of the Text Encoding Initiative Guidelines, TEI P5 (TEI Consortium 2007).

The resources are freely available for research and include uniformly encoded basic language resources for a large number of languages. These mostly include languages for which resources are scarcer than those for English and the languages of Western Europe. Best covered are the Slavic languages, which are well known for their complex morphosyntax and MULTEXT-East is the first dataset that enables an empirical comparison between them on this level of description.

The MULTEXT-East resources have helped to advance the state-of-the-art in language technologies in a number of areas, e.g. part-of-speech tagging (Tufiş 1999; Hajič 2000), learning of lemmatisation rules (Erjavec and Džeroski 2004; Toutanova and Cherry 2009), word alignment (Tufiş 2002; Martin et al. 2005), and word sense disambiguation (Ide 2000). They have served as the basis on which to develop further language resources, e.g. the WordNets of the BalkaNet project (Tufiş et al. 2004) and the JOS linguistically tagged corpus of Slovene (Erjavec et al. 2010). The morphosyntactic specifications have become a de-facto standard for several of the languages, esp. Romanian, Slovene and Croatian, where large monolingual reference corpora are using the MULTEXT-East tagset in their annotation. The resources have also provided a model to which some languages still lacking publicly available basic language engineering resources (tagsets, lexica, annotated corpora) can link to, taking a well-trodden path. In this manner resources for several new languages have been added to the V4 resources.

[1] EAGLES-based harmonized tagsets have been also used for various other language resources, such as those of the LE-PAROLE project, which produced a multilingual corpus and associated lexica for 14 European languages (Zampolli 1997).

Table 1 MULTEXT-East resources by language and resource type

Language	Language family	MSD specifications	MSD lexicon	1984		
				MSD	s-Align	Struct
English	Germanic	X	X	X	X	X
Romanian	Romance	X	X	X	X	X
Polish	West Slavic	X	X	X	O	–
Czech	West Slavic	X	X	X	X	X
Slovak	West Slavic	X	X	X	O	–
Slovene	South West Slavic	X	X	X	X	X
Resian	South West Slavic	X	X	–	–	–
Croatian	South West Slavic	X	–	–	–	–
Serbian	South West Slavic	X	X	X	X	X
Russian	East Slavic	X	X	O	O	X
Ukrainian	East Slavic	X	X	–	–	–
Macedonian	South East Slavic	X	X	X	X	–
Bulgarian	South East Slavic	X	X	X	X	X
Persian	Indo–Iranian	X	X	X	–	–
Estonian	Finno–Ugric	X	X	X	X	X
Hungarian	Finno–Ugric	X	X	X	X	X

Table 1 summarises the MULTEXT-East language resources by language (similar languages are grouped together and the ordering is roughly west to east), and by resource type. The resources marked with X are a part of the V4 release, while those marked with O have already been produced, and will be made available in the next release. Each type of resources is discussed in the next section, while an overview of all the languages included is given is Sect. 3.

2 The MULTEXT-East resources by type

2.1 The morphosyntactic specifications

The morphosyntactic specifications define word-level features (attributes and their values) which reside on the interface between morphology and syntax. The specifications also give the mapping from feature-structures used to annotate word-forms to the set of morphosyntactic descriptions (MSDs), which are compact strings used in the morphosyntactic lexica and, very often, for corpus annotation. So, for example, the MSD `Ncndl`, used for Slovene, maps to the feature-structure `Noun`, `Type:common`, `Gender:euter`, `Number:dual`, `Case:locative`. In addition to the formal parts, the specifications also contain commentary, bibliography, etc.

The common part of the specification gives the 14 MULTEXT defined *categories*, which mostly correspond to parts-of-speech, with a few introduced for technical reasons. Each category has a dedicated table defining its attributes, their

values, and their mapping to the (common) MSD strings. For each attribute-value pair it also specifies the languages it is appropriate for.

The second main part of the specifications consists of the language-specific sections. These, in addition to the introductory matter, also contain sections for each category with their tables of attribute-value definitions. These tables are similar to the common tables in that they repeat the attributes and their values, although only those appropriate for the language. However, they can also re-specify the position of the attributes in the MSD string, leading to much shorter and more readable MSD tags for the language.

The tables can furthermore contain localisation information. This enables expressing the feature-structures and MSDs either in English, or in the language that is being described, making them more suitable for use by native speakers of the language. Finally, each language particular section contains an index with all its valid MSDs, thus specifying the MSD tagset for the language. This is an important piece of information, as a tagged corpus can then be automatically validated against this authority list, and the tagset can be statically transformed into various other formats.

The specifications come with associated XSLT stylesheets, which take the specifications as input, usually together with certain parameters, and produce either XML, HTML or text output, depending on the stylesheet. Three classes of transformations are provided. The first help in adding a new language to the specifications themselves, the second transform the specifications into HTML for reading, and the third transform (and validate) a list of MSDs. The outputs of the second and third class of transformations are included in the MULTEXT-East distribution.

2.2 The morphosyntactic lexica

The MULTEXT-East morphosyntactic lexica have a simple structure, where each lexical entry is composed of three fields: (1) the *word-form*, which is the inflected form of the word, as it appears in the text, except for sentence-initial capitalisation; (2) the *lemma*, the base-form of the word, which e.g. serves as the head-word in a dictionary; and (3) the MSD, i.e. the morphosyntactic description, according to the language particular specifications. This simple lexical format is used in a number applications, such as the popular TreeTagger (Schmid 1994).

The sizes of the lexica vary considerably: Slovak and Macedonian have roughly 80,000 lemmas, mapping to over 1,000,000 entries, the majority offer medium sized lexica in the range of 15–50,000 lemmas, and a few are smaller, with Persian only covering the lemmas of "1984" and Resian simply giving examples for each MSD. However, even with the smaller lexica it should be noted that they cover the most morphologically complex words, such as pronouns (for Slavic languages) and high frequency open class words, providing a good starting point for the development of more extensive lexical resources. Also, all the languages that have an annotated "1984" corpus contain the entries for all its word-forms in the lexicon, providing a link between the lexicon and corpus.

2.3 The "1984" corpus

The parallel MULTEXT-East corpus consists of the novel "1984" by G. Orwell and its translations. This corpus is small (about 100,000 tokens per language), esp. by today's standards, and consists of only one text. Nevertheless, it provides an interesting experimentation dataset, as there are still very few uniformly annotated many-way parallel corpora.

The corpus is available in a format (given as "1984 struct" in Table 1) where the novel is extensively annotated for structures which would be mostly useful in the context of a digital library, such as verse, lists, notes, names, etc. More interestingly, the "1984" also exists as a separate corpus ("1984 MSD" in Table 1), which uses only basic structural tags but annotates each word with its context-disambiguated and—for most of the languages—hand-validated MSD and lemma. This dataset provides the final piece of the morphosyntactic triad, as it contextually validates the specifications and lexicon, and provides examples of actual usage of the MSDs and lexical items. It is useful for training part-of-speech taggers and lemmatisers, or for studies involving word-level syntactic information in a multilingual setting, such as advanced models of machine translation.

The "1984" corpus comes with separate alignment files (given as "1984 s-align" in Table 1) containing hand-validated sentence alignments between English and the translations, as well as pair-wise alignments between the other languages (so, currently, together 45 bi-lingual alignments) and a multi-way alignment spanning over all the 9 aligned languages.

3 MULTEXT-East by language

This section gives an overview of all the languages included in MULTEXT-East, concentrating on the origin of their MULTEXT-East resources and on publications that further detail their construction and use. Unless otherwise noted below, the linguistic annotation of the "1984" corpus has been, for the languages that have this corpus (c.f. Table 1), manually verified.

English is the hub language of the project: the English "1984" corpus is the source for the translations and the pivot for alignments, the English names of the morphosyntactic features serve as their canonical representation, the TEI element and attribute names are in English, as is the documentation of the resources. The English MULTEXT-East resources were already developed in the MULTEXT project, but were later adapted to be better harmonised with MULTEXT-East. However, the English MSD tagset has not really caught on and mappings to more widely used tagsets, such as those of CLAWS/BNC or the Penn TreeBank tagsets, have not been developed. Nevertheless, as discussed in the Introduction, the parallel "1984" with English as its hub has been used in many experiments.

Romanian resources were already part of the the original MULTEXT-East project results, and have not been substantially modified since. The specifications then served as the basis for various Romanian morphological lexica and annotated corpora and have become a de-facto standard for morphosyntactic annotation of the

language. The team led by Dan Tufiş has also published on a large number of experiments that used the resources as their dataset, esp. part-of-speech tagging (Tufiş 1999) and word alignment (Tufiş 2002).

Polish was added in Version 4, and Kotsyba et al. (2009) gives a detailed account of the theoretical background, the resources employed and the process of integrating the Polish language into MULTEXT-East. The morphosyntactic specifications are based on the flexemic tagset for Polish (Przepiórkowski and Woliński 2003), used e.g. for the annotation of the IPI PAN corpus of Polish (Przepiórkowski 2006), and this corpus was also taken as the source for constructing the MULTEXT-East lexicon. The tagging of "1984" was performed automatically, with the help of TaKIPI program (Piasecki 2007), developed for tagging Polish using the IPI PAN tagset.

Czech resources were produced as part of the original MULTEXT-East project and have not been substantially modified since. The morphosyntactic specifications essentially define a subset of the specifications for Czech described in Hajič (2002). The tagset developed by Hajič et al. is nowadays used as a standard for morphosyntactic annotations of the majority of Czech corpora, so the MULTEXT-East specifications have not been used outside of the project.

Slovak was added to the MULTEXT-East resources in Version 4 (Garabík et al. 2009). The morphosyntactic specifications were designed taking into account the tagset used in the Slovak National Corpus (Horák et al. 2004). Slovak has one of the largest MULTEXT-East lexica, with over 75,000 lemmas and almost 2 million entries. There is an automatic conversion software to convert the Slovak National Corpus tagset into MULTEXT-East MSDs, which was used in the construction of the lexicon and in the annotation of the "1984" corpus, with Garabík and Gianitsová-Ološtiaková (2005) giving the details of the annotation procedure.

Slovene has a special status in the context of MULTEXT-East, because it served as the testing ground for modifications in the overall structure of the resources. The first version of the Slovene specifications and lexicon was produced in the scope of the MULTEXT-East project and were based on the large morphological lexicon by the Slovene HLT company Amebis. The original specifications were subsequently modified for use in the 100 million word Slovene reference corpus FIDA (Krek et al. 1998). Since then the specifications have been used in a number of other corpus projects, most notably the 600 million word FidaPLUS reference corpus of Slovene (Špela Arhar and Gorjanc 2007). In the scope of the Slovene JOS project, which had the goal of producing freely available tagged corpora of Slovene (Erjavec et al. 2010), the specifications were substantially modified, taking into account the experiences of using them for over 10 years. The JOS morphosyntactic specifications then became the MULTEXT-East Version 4 specifications for Slovene. The Slovene MULTEXT-East resources have been used in a number of projects: in addition to the already mentioned FIDA, FidaPLUS and JOS corpora, they were also used e.g. as the basis for the first treebank of Slovene (Džeroski et al. 2006), included in the 2006 CoNLL-X shared task on multi-lingual dependency parsing (Buchholz and Marsi 2006).

Resian is a very distinct dialect of Slovene spoken in the Resia valley in north-eastern Italy, close to the border with Slovenia. Because of its remote location

outside of Slovenia, the dialect has phonetical and morphological properties that are very different from standard Slovene, and from most other Slovene dialects (Steenwijk 1992). The Resian specifications were added to MULTEXT-East in Version 3 by Han Steenwijk from the University of Padova and then served as the basis for developing a basic lexicon and annotated corpus of Resian, available at http://www.resianica.it/.

Croatian specifications were added in MULTEXT-East Version 2. These specifications have since become a de-facto standard for Croatian, as they were used both for the morphosyntactic tagging of the 100-million-word Croatian National Corpus (Tadić 2002) and in the Croatian Morphological Lexicon (Tadić 2003). Unfortunately, other than the morphosyntactic specifications, none of the other Croatian resources are accessible through MULTEXT-East.

Serbian resources were added to MULTEXT-East in Version 3 (Krstev et al. 2004) and the lexicon has been substantially enlarged for Version 4. The morphosyntactic specifications are based on the feature specifications as used in the Serbian morphological lexicon (Vitas and Krstev 2001) developed in the INTEX/NooJ finite-state toolbox (Silberztein 1999). This lexicon has been automatically converted into the MULTEXT-East format and included in the MULTEXT-East resources.

Russian "1984" as a structurally annotated document with alignments was already available in Version 2 of the resources, however, the specifications and the lexicon have been added only in Version 4 (Sharoff et al. 2008). The developed specifications, MSD tagset and lexicon took as the basis the Russian National Corpus (Sharoff 2005), which is comparable to the BNC Sampler in its size and accuracy of annotation, and HANCO (Kopotev and Mustajoki 2003), developed at the University of Helsinki. An automatically tagged corpus with the MULTEXT-East tagset, as well as tagging models for various taggers are freely available from http://corpus.leeds.ac.uk/mocky/.

Ukrainian was added in Version 4 (Derzhanski and Kotsyba 2009). The specifications and the lexicon are based on the Ukrainian Grammatical Dictionary (UGD) developed at the Ukrainian Academy of Sciences by Igor V. Shevchenko, and the morphological analyzer UGTag, which uses an extended version of the UGD. The MULTEXT-East Ukrainian lexicon constitutes the first publicly available lexicon for the language.

Macedonian was also added in Version 4. The morphosyntactic specifications were developed from scratch and the lexicon was converted from a previously available INTEX lexicon (Petrovski 2004). The INTEX finite-state toolkit allows for specifications of morphological patterns and the Macedonian lexicon contains not only the full inflectional paradigms of the lemmas but also (the inflectional patterns of) automatically computed derivational variants of the base lemmas, in particular about 10,000 adjectives, derived from verbs (Zdravkova and Petrovski 2007). This makes it, in terms of the number of lemmas (over 80,000), the largest lexicon of all languages covered. The "1984" corpus was also developed and sentence aligned with English. The corpus is currently annotated only with non-disambiguated MSDs and lemmas—Macedonian does not, as yet, have a manually tagged corpus. This also means that the encoding of the annotated corpus is

somewhat different from the others, as it needs to represent the ambiguity in the lemma and MSD assignment to the tokens. The developed Macedonian resources have been used in several experiments in tagger and lemmatiser induction (Vojnovski et al. 2005; Ivanovska et al. 2005), and a description of their development and potential use for machine translation experiments is given in Stolić and Zdravkova (2010).

Bulgarian resources were part of the original MULTEXT-East project. The language already had various morphosyntactic lexica using different specifications at the start of the MULTEXT-East project, and the MULTEXT-East specifications were a derivation of one of them, Slovnik. A detailed comparison of the tagsets (including the EAGLES one) is given in Slavcheva (1997). The Slovnik tagset was later adapted for the purposes of the BulTreeBank project (Simov et al. 2002). Although some plans were made to update the morphosyntactic specifications for Version 4 (Dimitrova and Rashkov 2009; Garabík et al. 2009), they have not been put into practice, so the specifications and lexicon have not changed since the initial release. The annotated corpus was also only automatically tagged, with the tagset being a reduction of the MSDs defined in the specifications.

Persian (Farsi) resources were developed by QasemiZadeh and Rahimi (2006) and were added to MULTEXT-East in Version 4. The specifications were written from scratch taking into account mainly standard grammars of Persian. The lexicon and annotated "1984" also become available via ELDA in 2010.

Estonian resources were part of the original MULTEXT-East project (Dimitrova et al. 1998) and have not changed since. They have also not been directly used in any further work on Estonian.

Hungarian resources were also part of the original MULTEXT-East project, although the specifications were significantly revised for Version 4. The original specifications and lexicon were based on the encoding already used for Hungarian, which uses a feature-structure mechanism to represent morphosyntactic information in lexica (Prószéky 1995; Prószéky and Kis 1999). This system is still the prevalent one in use for tagging Hungarian texts. However, a manually annotated corpus which does use (a modified form of) the Hungarian MULTEXT-East specifications was developed by Alexin et al. (2003), primarily to serve as a gold standard for the development of morphosyntactic tagging programs, and as the basis for a Hungarian treebank.

4 Conclusions

The resources described in the paper are distributed on the Web, from the URL http://nl.ijs.si/ME/. The morphosyntactic specifications and documentation are freely available. For the lexica and the corpus the user has to fill out a Web-based agreement form restricting the use of resources for research. In the future we plan to include the resources in some other repositories of language resources as well.

Further work on the resources could proceed in a number of directions. The MULTEXT-East morphosyntactic specifications currently lack consistency between the languages (Przepiórkowski and Woliński 2003; Derzhanski and Kotsyba 2009;

Feldman and Hana 2010), and a typology of the problems is summarised in Rosen (2010). Problematic cases are divided into those where different features in various languages are used to describe the same phenomenon, and, conversely, the same features are used to describe different phenomena. Furthermore, certain tags are too specific and hard to extend to cover similar phenomena in another language. Some steps in harmonising the MULTEXT-East specifications have already been undertaken in the context of converting them into an OWL DL ontology (Chiarcos and Erjavec 2011), which enables logical inferences over feature sets to be made on the basis of partial information. This process also pin-pointed inconsistencies, which could then be, to an extent, resolved in the context of the ontology. The next step in the development of the specifications and associated tagsets, currently under development, is to link them to universal vocabularies, such as the isoCat Data Category Registry (Kemps-Snijders et al. 2008) and GOLD, the General Ontology for Linguistic Description (Farrar and Langendoen 2003).

Given that the specifications are grounded in the parallel corpus, it would also be interesting to explore machine-translation based (semi)automatic mapping procedures between MSDs and feature bundles for the languages. Such research would also be illuminating from a comparative linguistics point of view.

Finally, we could continue to add new languages to the MULTEXT-East resources. The most interesting ones are the missing languages from Eastern and Central Europe, in particular Lithuanian and Latvian, where some initial work has already been done. It would, of course, also be nice to integrate the MULTEXT (-West) resources into the -East off-shoot.

Acknowledgments The author would like to thank Radovan Garabik, Natalia Kotsyba, Katerina Zdravkova, and Darja Fišer for their helpful comments and suggestions. Work on the MULTEXT-East resources was initially supported by the EU project MULTEXT-East "Multilingual Text Tools and Corpora for Central and Eastern European Languages", the US NSF grant IRI-9413451 and the EU Concerted Action TELRI "Trans-European Language Resources Infrastructure". Work on the second release was supported by the EU Project CONCEDE "Consortium for Central European Dictionary Encoding", while the work on the third release was partially funded by a the NEH grant to the TEI Task Force "SGML–XML migration". Work on the fourth release was supported by the EU project MONDILEX "Conceptual Modeling of Networking of Centres for High-Quality Research in Slavic Lexicography and their Digital Resources". The work on the resources has been additionally supported by bi-lateral projects between Slovenia and Serbia, Slovenia and Macedonia, as well as individual partners' grants and contracts.

References

Alexin, Z., Gyimóthy, T., Hatvani, C., Tihanyi, L., Csirik, J., Bibok, K., et al. (2003). Manually annotated hungarian corpus. In *Proceedings of the tenth conference on European chapter of the association for computational linguistics* (*EACL'03*) (pp. 53–56).

Arhar, Š., & Gorjanc, V. (2007). Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa (the FidaPLUS corpus: A new generation of the Slovene reference corpus). *Jezik in slovstvo*, *52*(2), 95–110.

Buchholz, S., & Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning* (*CoNLL-X*) (pp. 149–164). Morristown, NJ, USA: ACL.

Chiarcos, C., & Erjavec, T. (2011) OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. In *Proceedings of the 5th linguistics annotation workshop* (*LAW-V*), ACL.

Derzhanski, I. A., & Kotsyba, N. (2009). Towards a consistent morphological tagset for Slavic languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. In *Proceedings of the Mondilex third open workshop: Metalanguage and encoding scheme design for digital lexicography* (pp. 9–26). Bratislava, Slovakia: Ľ. Štúr Institute of Linguistic, Slovak Academy of Sciences.

Dimitrova, L., & Rashkov, P. (2009). A new version for Bulgarian MTE morphosyntactic specifications for some verbal forms. In *Proceedings of the Mondilex second open workshop: Organization and development of digital lexical eesources* (pp. 30–37). Kyiv, Ukraine: Dovira Publishing House.

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H. J., Petkevič, V., & Tufiş, D. (1998). MULTEXT-East: Parallel and comparable corpora and lexicons for six Central and Eastern European languages. In *Proceedings of the COLING-ACL'98* (pp. 315–319). Montréal, QC, Canada: ACL.

Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtsky, Z., & Žele, A. (2006). *Towards a Slovene dependency treebank. In Proceedings of the fifth international conference on language resources and evaluation* (*LREC'06*), Genoa.

EAGLES. (1996). *Expert advisory group on language engineering standards*. http://www.ilc.pi.cnr.it/EAGLES/home.html.

Erjavec, T. (2004). MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of the fourth international conference on language resources and evaluation* (*LREC'06*), Lisbon.

Erjavec, T. (2010) MULTEXT-East version 4: Multilingual morphosyntactic specifications, lexicons and Corpora. In *Proceedings of the seventh international conference on language resources and evaluation* (*LREC'06*), Valetta.

Erjavec, T., & Džeroski, S. (2004). Machine learning of language structure: Lemmatising unknown Slovene words. *Applied Artificial Intelligence*, *18*(1), 17–41.

Erjavec, T., Fišer, D., Krek, S., & Ledinek, N. (2010). *The JOS linguistically tagged corpus of Slovene. In Proceedings of the seventh international conference on language resources and evaluation* (*LREC'10*), Valetta.

Farrar, S., & Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT International, 7*(3), 97–100.

Feldman, A., & Hana, J. (2010). *A resource-light approach to morpho–syntactic tagging. Language and computers: Studies in practical linguistics* (Vol. 70). Amsterdam: Rodopi.

Garabík, R., & Gianitsová-Ološtiaková, L. (2005). Manual morphological annotation of the Slovak translation of Orwell's novel 1984: Methods and findings. In *Proceedings of the Slovko conference* "*computer treatment of Slavic and East European languages*". Bratislava: Veda.

Garabík, R., Majchráková, D., & Dimitrova, L. (2009). Comparing Bulgarian and Slovak MULTEXT-East morphology tagset. In *Proceedings of the Mondilex second open workshop: Organization and development of digital lexical resources* (pp. 38–46). Kyiv, Ukraine: Dovira Publishing House.

Hajič, J. (2000). Morphological tagging: Data versus dictionaries. In *Proceedings of the ANLP/NAACL 2000* (pp. 94–101). Seattle.

Hajič, J. (2002). *Disambiguation of rich inflection* (*computational morphology of Czech*) (Vol. 1). Prague: Karolinum Charles University Press.

Horák, A., Gianitsová, L., Šimková, M., Šmotlák, M., & Garabík, R. (2004). Slovak national corpus. In *Proceedings of the text speech and dialogue conference* (*TSD'04*), Brno.

Ide, N. (1998). Corpus encoding standard: SGML guidelines for encoding linguistic corpora. In *Proceedings of the first international conference on language resources and evaluation* (*LREC'98*) (pp. 463–470). Granada.

Ide, N. (2000). Cross-lingual sense determination: Can it work? *Computers and the Humanities, 34*, 223–234.

Ide, N., & Véronis, J. (1994). Multext (multilingual tools and corpora). In *Proceedings of the 15th international conference on computational linguistics* (*CoLing'94*) (pp. 90–96). Kyoto.

Ivanovska, A., Zdravkova, K., Džeroski, S., & Erjavec, T. (2005). Learning rules for morphological analysis and synthesis of Macedonian nouns. In *Proceedings of the 8th international conference information society, IS 2005*. Ljubljana: Jožef Stefan Institute.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. E. (2008). ISOcat: Corralling data categories in the wild. In *Proceedings of the sixth international conference on language resources and evaluation* (*LREC'08*), Marrakech.

Kopotev, M., & Mustajoki, A. (2003) *Principy sozdanija Hel'sinkskogo annotirovannogo korpusa russkih tekstov (HANCO) v seti internet. Naučno-tehničeskaja informacija* (Ser. 2, pp. 33–37) (in Russian).

Kotsyba, N., Radziszewski, A., & Derzhanski, I. (2009). Integrating the Polish language into the MULTEXT-East family. In *Proceedings of the Mondilex fifth open workshop*: *Research infrastructure for digital lexicography*. Ljubljana, Slovenia: Jožef Stefan Institute.

Krek, S., Stabej, M., Gorjanc, V., Erjavec, T., Romih, M., & Holozan, P. (1998) *FIDA*: *A corpus of the Slovene language*. http://www.fida.net/.

Krstev, C., Vitas, D., & Erjavec, T. (2004). MULTEXT-East resources for Serbian. In *Proceedings B of the 7th international multiconference information society*: *Language technologies* (pp. 108–114). Ljubljana: Jožef Stefan Institutue.

Martin, J., Mihalcea, R., & Pedersen, T. (2005). Word alignment for languages with scarce resources. In *Proceedings of the ACL workshop on building and using parallel texts* (pp. 65–74). Ann Arbor.

Petrovski, A. (2004). Morphological processing of nouns in Macedonian language. In *Proceedings of the 7th intex/nooj workshop*, Tours.

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly, 11*, 151–167.

Prószéky, G. (1995). Humor: A morphological system for corpus analysis. In *Proceedings of the first European TELRI seminar*: *Language resources for language technology* (pp. 149–158). Tihany, Hungary.

Prószéky, G., & Kis, B. (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th ACL, association for computational linguistics* (pp. 261–268).

Przepiórkowski, A. (2006). The potential of the IPI PAN corpus. *Poznań Studies in Contemporary Linguistics, 41*, 31–48.

Przepiórkowski, A., & Woliński, M. (2003). A flexemic tagset for Polish. In *Proceedings of the EACL workshop on morphological processing of Slavic languages*. ACL.

QasemiZadeh, B., & Rahimi, S. (2006) Persian in MULTEXT-East framework. In *Proceedings of the 5th international conference on natural language processing* (*FinTAL'06*) (pp. 541–551). Turku, Finland.

Rosen, A. (2010). Morphological tags in parallel corpora. In F. Čermák, A. Klégr, & P. Corness (Eds.), *InterCorp*: *Exploring a Multilingual corpus*. Praha: Nakladatelství Lidové noviny.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing* (pp. 44–49).

Sharoff, S. (2005). Methods and tools for development of the Russian reference corpus. In D. Archer, A. Wilson, & P. Rayson (Eds.), *Corpus linguistics around the world* (pp. 167–180). Amsterdam: Rodopi.

Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., & Divjak, D. (2008). Designing and evaluating a Russian tagset. In *Proceedings of the sixth international conference on language resources and evaluation* (*LREC'08*). Marrakech.

Silberztein, M. (1999). Text Indexing with INTEX. In: Computers and the humanities (vol. 33(3)). Kluwer Academic Publishers.

Simov, K., Popova, G., & Osenova, P. (2002). HPSG-based syntactic treebank of Bulgarian (BulTreeBank). In A. Wilson, P. Rayson, & T. McEnery (Eds.), *A rainbow of corpora: Corpus linguistics and the languages of the world* (pp. 135–142). Munich: Lincom-Europa.

Slavcheva, M. (1997). *A comparative representation of two Bulgarian morphosyntactic tagsets and the EAGLES encoding standard*. Technical Report TELRI (Trans European Language Resources Infrastructure).

Sperberg-McQueen, C. M., & Burnard, L. (Eds.). (1994). *Guidelines for electronic text encoding and interchange P3*. Chicago and Oxford: Association for Computers and the Humanities/Association for Computational Linguistics/Association for Literary and Linguistic Computing.

Steenwijk, H. (1992). *The Slovene Dialect of Resia San Giorgio*. Amsterdam-Atlanta: Rodopi.

Stolić, M., & Zdravkova, K. (2010). Resources for machine translation of the Macedonian language. In *Proceedings of the ICT innovations conference*, Ohrid.

Tadić, M. (2002). Building the Croatian national corpus. In *Proceedings of the third international conference on language resources and evaluation* (*LREC'02*) (pp. 441–446). Las Palmas.

Tadić, M. (2003). Building the Croatian morphological lexicon. In *Proceedings of the EACL workshop on morphological processing of Slavic languages*, ACL.

TEI Consortium. (2007). *TEI P5: Guidelines for electronic text encoding and interchange*. TEI Consortium, URL: http://www.tei-c.org/Guidelines/P5/.

Toutanova, K., & Cherry, C. (2009). A global model for joint lemmatization and part-of-speech prediction. In *Proceedings of the 47th annual meeting of the ACL (ACL'09)* (pp. 486–494). Singapore.

Tufiş, D. (1999). Tiered tagging and combined language model classifiers. In F. Jelinek & E. Noth (Eds.), *Text, speech and dialogue no. 1692 in lecture notes in artificial intelligence* (pp. 28–33). Berlin: Springer.

Tufiş, D. (2002). A cheap and fast way to build useful translation lexicons. In *Proceedings of the 19th annual meeting of the ACL (ACL'02)*. Association for Computational Linguistics.

Tufiş, D., Cristea, D., & Stamou, S. (2004). BalkaNet: Aims, methods, results and perspectives: A general overview. *Romanian Journal of Information Science and Technology*,*7*(1–2), 9–43.

Vitas, D., & Krstev, C. (2001). Intex and slavonic morphology. In *4es Journées INTEX*. Bordeaux.

Vojnovski, V., Džeroski, S., & Erjavec, T. (2005). Learning PoS tagging from a tagged Macedonian text corpus. In *Proceedings of the 8th international conference information society, IS 2005*. Ljubljana: Jožef Stefan Institute.

Zampolli, A. (1997). The PAROLE project. In *Proceedings of the second European TELRI seminar*: *Language applications for multilingual Europe* (pp. 185–210). Kaunas, Lithuania.

Zdravkova, K., & Petrovski, A. (2007). Derivation of Macedonian verbal adjectives. In *Proceedings of international conference "recent advances in natural language processing"* (*RANLP'07*) (pp. 661–665).