04 MULTEXT-East Morphosyntax

The MULTEXT-East framework for morphosyntactic annotation of text corpora defines character codes, referred to as MSD-tags (with 'MSD' standing for morphosyntactic description). For example, the "Ncmsn" tag represents a set of grammatical features "Noun Type=common Gender=masculine Number=singular Case=nominative". This annotation system has been established for 20 languages or dialects, including all Slavic languages.
The use of MULTEXT-East tags for Slovene began in 1996 and has since continued in all subsequent open corpora of Slovene, whether manually or automatically annotated, up until the emergence of the Universal Dependencies morphosyntactic annotation framework, which is now gradually taking over the role that MULTEXT-East played for decades.

Introduction to Tags

In this chapter, we outline the design of the MULTEXT-East specifications.

The multilingual MULTEXT-East specifications are written in XML, following the TEI recommendations, and define the morphosyntactic features (attributes and their values) of words, i.e. the features of words that lie at the intersection of morphology and syntax. The specifications also provide a mapping of sets of these features to morphosyntactic descriptions (MSDs), which are compact strings used in corpus annotation. For example, the MSD "Ncndl" is mapped to the features "Noun, Type:common, Gender:neuter, Number:dual, Case:locative". In addition to the formal parts, the specifications also contain comments, bibliography, etc.

The common part of the specifications delineates 14 established MULTEXT categories, mostly corresponding to parts of speech, although some categories have been introduced for technical reasons. Each category features a dedicated table outlining its attributes, their values and their mapping to MSD strings. For each attribute-value pair, it also defines the languages to which that pair applies.

The second main part of the specifications includes language-specific segments for each language within the framework. In addition to the introduction, these segments offer detailed tables for each category, providing attribute value definitions. While these tables mirror the general tables in presenting attributes and their values, they are tailored to only include those that apply to the language at hand. Furthermore, these language-specific tables can also redefine the position of attributes in the MSD string, leading to significantly more concise and more readable MSD tags for that language.

Language-specific tables can include localization information as well. This feature enables the representation of attributes and MSDs either in English or in the language being described, making them more suitable for use by native speakers of the language. Finally, the language-specific section also lists all valid MSDs, thus defining the set of MSD tags for that language. The set of allowed MSDs is an important resource: it not only allows for automated verification of corpora tagged with these identifiers but also facilitates the conversion of the tag set into diverse formats.

The XML or TEI specifications come with accompanying XSLT programs designed to process the specifications as input (with optional parameters), generating outputs in XML, HTML, or plain text, depending on the chosen stylesheet. Three classes of transformation are offered: The first assists in incorporating a new language into the specifications; the second converts the specifications into HTML format for ease of reading; and the third processes (and validates) a list of MSDs. Outputs from the latter two classes of transformation are provided as part of the MULTEXT-East distribution package.

Annotation Guidelines

This chapter summarizes the annotation guidelines for the MULTEXT-East morphosyntax as applied to Slovene texts. The guidelines are arranged from the latest, up-to-date version to the oldest version.

Version 2.0 (25-02-2023)

Project Development of Slovene in a Digital Environment

HOLOZAN, Peter, KREK, Simon, PIVEC, Matej, RIGAČ, Simon, ROZMAN, Simon, VELUŠČEK, Aleš, PORI, Eva, ARHAR HOLDT, Špela, 2023: *Specifikacije za učni korpus: lematizacija in MSD.* Različica 2.0. Rezultat projekta Razvoj slovenščine v digitalnem okolju.

[DOCX] [PDF] - only in Slovene

Version 1.0 for non-standard Slovene (21-12-2016)

Project Janes - Resources, Tools and Methods for the Research of Nonstandard Internet Slovene

ČIBEJ, Jaka, ARHAR HOLDT, Špela, ERJAVEC, Tomaž, FIŠER, Darja, ZUPAN, Katja, 2016: *Smernice za označevanje računalniško posredovane komunikacije: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladenjsko označevanje.* Različica 1.0. Rezultat projekta Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine. [PDF] - only in Slovene

Version 1.0 (2008)

Project Communication in Slovene

HOLOZAN, Peter, KREK, Simon, PIVEC, Matej, RIGAČ, Simon, ROZMAN, Simon, VELUŠČEK, Aleš, 2008: *Specifikacije za učni korpus.* Različica 1.0. Kazalnik K2 projekta Sporazumevanje v slovenskem jeziku. [PDF] - only in Slovene

Specifications MULTEXT-East:

Specifications MULTEXT-East V6 on GitHub: https://github.com/clarinsi/mte-msd

Specifications MULTEXT-East V6 in TEI: https://nl.ijs.si/ME/V6/msd/xml/

Specifications MULTEXT-East V6 for reading: https://nl.ijs.si/ME/V6/msd/html/index.html

Specifications MULTEXT-East V6 for Slovene:

- in TEI: https://nl.ijs.si/ME/V6/msd/xml/msd-sl.spc.xml
- for reading: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html
- tags in table format: https://nl.ijs.si/ME/V6/msd/tables/msd-human-sl.tbl
- tags encoded in TEI: https://nl.ijs.si/ME/V6/msd/tables/msd-fslib2-sl.xml

References and Links

This chapter compiles relevant references and provides links to projects where the MULTEXT-East morphosyntax has been developed and applied to Slovene texts.

Projects, in which the system has been developed or applied

MULTEXT-East - Multilingual corpora and text tools for Central and East European langauges

JOS - Linguistic Annotation of Slovene: Methods and Resources

Communication in Slovene

Janes - Resources, Tools and Methods for the Research of Nonstandard Internet Slovene

Development of Slovene in a Digital Environment

Training corpora containing manually revised MULTEXT-East tags

Krek, Simon; et al., 2019, Training corpus ssj500k 2.2, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1210.

Arhar Holdt, Špela; et al., 2022, Training corpus SUK 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1747.

Lenardič, Jakob; et al., 2022, CMC training corpus Janes-Tag 3.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1732.

References

Erjavec, Tomaž; Fišer, Darja; Krek, Simon in Ledinek, Nina. 2010. The JOS Linguistically Tagged Corpus of Slovene. V: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valeta, Malta, Maj. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf [PDF]

Erjavec, Tomaž. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. Language Resources and Evaluation, 46(1): 131–142. DOI: 10.1007/s10579-011-9174-8 [PDF]

Erjavec, Tomaž. 2017. MULTEXT-East. V (Nancy Ide, James Pustejovsky, ur.): Handbook of linguistic annotation. pp. 441-462. Springer. DOI: 10.1007/978-94-024-0881-2_17

Krek, Simon; Erjavec, Tomaž; Dobrovoljc, Kaja; Gantar, Polona; Arhar Holdt, Špela; Čibej, Jaka in Brank, Janez. The ssj500k training corpus for Slovene language processing. V: Fišer, D. in Erjavec, T. Jezikovne tehnologije in digitalna humanistika: zbornik konference: 24.-25. september 2020, Ljubljana, Slovenija. Ljubljana: Inštitut za novejšo zgodovino, 2020. Str. 24–33.

http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene-Language-Processing.pdf [PDF]