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.

Revision #1
Created 13 November 2023 13:25:18 by Tina Munda
Updated 13 November 2023 13:27:20 by Tina Munda