Parallel corpora within the Russian National Corpus: current state and the study of Russian functional words.

The parallel corpora within the Russian National Corpus has recently approached the size of 100 million tokens and feature bilingual pair of Russian with such typologically diverse languages as Armenian, Bashkir, Buryat, Chinese and different Slavic and non-Slavic European languages, as well as a multilingual section (cf. the notion of "massive parallel texts" according to Cysouw, Wälchli 2006). The talk presents the state of the art in morphological annotation and alignment of the texts in these languages, including the choice of tokenizers, transliteration, glossing rules, and elements of semantic annotation such as dictionary entries.

The case studies presented in the paper concern the models and stimuli of translations for Russian function words such as parentheses, discourse markers, and conjunctions of cause. Often they have no direct counterpart in the text in another language (such as *byvalo* 'used.to.be', a semi-grammaticalized marker of Past Habitual, cf. Dahl 1985) or are rendered by constructions with a more general meaning (such as *blago* 'because (of a positive factor)' translated by *while* or *as*). However in a fraction of cases they have correspondences that reflect different facets of their semantics; these correspondences contribute to a more thorough description of the Russian units. Such techniques of visualizing the results as NeighbourNet are used, showing the differences and similarities in the correspondences of the Russian items. The research addresses the boundaries between discourse and grammatical markers and different treatment of the same semantics across languages with different structure.

References

Cysouw, M., Wälchli B. (eds.). *Parallel Texts. Using Translational Equivalents in Linguistic Typology. Theme issue in Sprachtypologie & Universalienforschung STUF* 60.2, 2007.

Dahl, Ö. 1985. Tense and aspect systems. Oxford: Blackwell.