Olga Lyashevskaya. Pavia, September 17-21, 2012.

Syllabus

This course will cover the theoretical approaches to designing and compiling corpora for various purposes as well as practicalities such as text preprocessing, classification, large-scale linguistic annotation, and multiword tagging. Special focus will be given to the Russian National Corpus, and you will have the opportunity to develop the skills required to build the corpus of your dream. We will also learn how to use semantic and morphological information in quantitative corpus-based research.

The course is structured in four sessions.

L1. Measures of collocations and multiword expressions

Key words: n-grams, collocations, colligations, collostructions, multiword expressions/extended lexical units, idioms/phrasal constructions/frozen sentences, prefabs, compositionality. Collocations dictionaries and other lexical sources. Extraction of collocations from a corpus: statistical association measures. Patterns and stop words. Word forms vs. lemmata. Distinctiveness lists: contrasting registers. Evaluation of the MWE lists.

*Greaves, C. and M. Warren. 2010. 'What can a corpus tell us about multi-word units?'. In A. O'Keeffe and M. McCarthy (eds.), The Routledge handbook of Corpus Linguistics. London: Routledge, pp. 212-226.

Stefan Evert's page on Computational Approaches to Collocations at http://www.collocations.de

Manning, Christopher D. and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press. Chapter 5. http://nlp.stanford.edu/fsnlp/promo/colloc.pdf

L2. Criteria for compiling a national corpus: the Russian National Corpus

Key words: corpus size, time span, domain and genre balance, representativeness. BNC as a model of a national corpus. National corpora around the world and greater corpora. General-purpose corpora vs. specialized corpora. Written, spoken and multimodal data. Text balance and practicalities. Copyright matters. Text collection and preprocessing. Layers of annotation. Metatextual information. Manual and automatic tagging. Language and culture-specific challenges.

*O'Keeffe, Anne and Michael McCarthy (eds.). 2010. The Routledge Handbook of Corpus Linguistics. London: Routledge. Section 2: Building and Designing a Corpus: What are the Key Considerations?, pp.

Sharoff, Serge. 2005. Methods and tools for development of the Russian Reference Corpus. In: D. Archer, A. Wilson, P. Rayson (eds.), Corpus Linguistics around the World. Amsterdam: Rodopi, pp. 167-180.

L3. Semantic and morphological annotation

Key words: tagset; lexico-semantic groups, ontologies, WordNet, frames, TimeML, SpaceML, anaphoric relations; tokens and types, word forms and lemmata, inflectional paradigm, morphosyntactic categories. Tokenization. PoS-annotation. Lemmatization. Grammatical tagsets, core grammatical features and extras. Grammatical dictionaries. Hypotheses for word forms not in a dictionary. Lemma and tag disambiguation. Semantic annotation. Word-sense disambiguation.

*Kustova Galina I., Olga N. Lashevskaja, Elena V.Paducheva, and Ekaterina V. Rakhilina. 2009. 'Verb taxonomy: from theoretical lexical semantics to practice of corpus tagging'. In: B. Lewandowska, K. Dziwirek (eds.), Cognitive Corpus Linguistics Studies. Frankfurt: Peter Lang, 2009.

Lashevskaja, Olga and Olga Mitrofanova. 2009. ‘Disambiguation of taxonomy markers in context: Russian nouns’. 17th Nordic Conference on Computational Linguistics (NODALIDA 2009). Odense, Denmark, May 14-16, 2009. NEALT Proceedings Series, vol. 4. P. 111-117.

Sharoff, S., M. Kopotev, T. Erjavec, A. Feldman, and D. Divjak. 2008. ‘Designing and evaluating a Russian tagset’. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08), Marrakech, Morocco. http://pages.cs.brandeis.edu/~marc/misc/proceedings/lrec-2008/pdf/78_paper.pdf

L4. Quantitative methods in cognitive linguistics

Some key concepts of cognitive linguistics through the lab. Basic techniques for collecting and analyzing empirical data using corpora and accessible statistical software.

Janda, Laura A. and Olga Lyashevskaya. 2011. ‘Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian’, Cognitive Linguistics, 22(4), pp. 719-763.

Sokolova, Svetlana, Olga Lyashevskaya, and Laura Janda (forthcoming). ‘The Locative Alternation and the Russian ‘empty’ prefixes: A case study of the verb *gruzit’* ‘load’’. In: Divjak, D. & St.Th. Gries (eds.). Frequency effects in language: linguistic representations. Mouton.