

The Helsinki Annotated Corpus of Russian Texts HANCO:  
a tool for teaching and learning of Russian

The HANCO Corpus project has been running since 2001 in the Department of Slavonic and Baltic Languages and Literatures at the University of Helsinki. The main principles of creation are:

- **Orientation to a wider audience.** In drawing up the HANCO corpus and its computer interface, the creators have kept in mind as potential users not only a narrow circle of experts, but also students and teachers of Russian, including L2 teachers. This does not mean that the use of linguistic terms is avoided, but the choice of parameters for a search is carried out in such a way as to minimize the amount of specialized knowledge required.
- **Orientation to the traditional conception** of a language. In compiling the set of parameters it has been preferred to establish theoretical concepts, which are used in widely known linguistic works and/or in textbooks on Russian grammar. This is due to the demand of setting a low threshold for the use of the Corpus.
- **Orientation to the accuracy** of the grammatical description, not to the amount of annotated material. The purpose is to create an annotated corpus containing more exact grammatical and functional information than is the case in existing or planned larger Russian corpora.
- **Orientation to multilevel grammatical information.** The HANCO corpus contains multilateral grammatical information including morphological, syntactic, and functional (semantic) characteristics. They can be combined in the process of searching.
- **Possibility of alternative interpretations.** Any researcher working with concrete language material will have come across the fact that it is difficult or even impossible in many cases to classify linguistic units in an unequivocal way. In creating the HANCO, we made the decision to accept the possibility of alternative interpretations of linguistic facts. Such seeming illegibility demands a lot of manual work, but it facilitates the searching of necessary information by the potential user.

The following types of linguistic information are to be included in the HANCO.

- **Morphological information.** In the HANCO, complete morphological description of every running word is given. The morphological analysis and the subsequent disambiguation procedure have been carried out automatically, with further manual processing.
- **Syntactic information.** This is given at two levels: clauses and sentences. The full description of units for every level is carried out according to the widely accepted tradition to the description of Russian syntax.
- **Functional-semantic information.** The Corpus is a component of the Contrastive Functional Syntax project conducted by Professor A. Mustajoki. Due to this link, the HANCO will also provide information based on semantic categories, the list of which is being elaborated by A. Mustajoki's scientific team.

The HANCO is created step by step. The first results (the morphological and syntactic parts) have already been achieved at [www.slav.helsinki.fi/hanco](http://www.slav.helsinki.fi/hanco), thus it can be used as a tool in teaching and learning of Russian. The report will present some possible applications of the corpus.