Corpus Instruments for Russian Grammar Studies


Corpus linguistics can be broadly defined in terms of two partially overlapping research dimensions. On the one hand, corpus linguistics is knowledge of how to compile and annotate linguistic corpora. On the other hand, corpus linguistics is a family of qualitative and quantitative methods of language study based on corpus data. The book presents the first steps taken by Russian corpus linguistics toward the development of language corpora and corpus-based resources as well as their use in grammatical and lexical analysis.

The first part of the book focuses on the annotation of Russian texts at several levels: lemmas, part of speech and inflectional forms, word formation, lexical-semantic classes, syntactic dependencies, semantic roles, frames, and lexical constructions. We discuss various theoretical principles and practical considerations motivating the corpus markup design, provide details on the creation of lexical resources (electronic dictionaries and databases) and text processing software, and consider complicated cases that present challenges for the annotation of corpora both manually and automatically. In most cases we describe the annotation of the Russian National Corpus (RNC, ruseorpora.ru) and its affiliate project FrameBank (framebank.ru).

Frequency data depend not only on the representativeness and balance of texts in a corpus, but also on the rules and tools used for annotation. The book addresses the development of evaluation standards for Russian NLP resources, namely, morphological taggers and dependency parsers. In addition, the book presents several experiments on automatic annotation and disambiguation: lemmatization of word forms not in the dictionary; word sense disambiguation based on vectors formed by lexical, semantic and grammatical cues of context; and semantic role labeling.

The final chapters of the first part of the book outline two types of frequency dictionaries based on the RNC data: a general-purpose frequency dictionary and a lexicogrammatical one.

The second part of the book presents an analysis of corpus data and includes a number of case studies of Russian grammar and lexical-grammatical interaction using quantitative methods. The key concept underlying our analysis is the behavioral profile (Hanks 1996; Divjak, Gries 2006), which is the frequency distribution of variable elements in
a linguistic unit as attested in a corpus. This covers grammatical profiles (the frequency
distribution of inflected forms of a word), constructional profiles (the frequency distri-
bution of argument or any other constructions attested for a key predicate), lexical and
semantic profiles (the frequency distribution of words and lexical-semantic classes in
construction slots or, more generally, in the context of a word), and radial category pro-
files (the frequency distribution of word senses and word uses across the radial category
network of a polysemous unit). We use grammatical, constructional, semantic, and radial
category profiling to study tense, aspect and mood specialization of Russian verb forms;
to identify singular-oriented and plural-oriented nouns; to investigate factors for prefix
choice and prefix variation in natural perfectives (chistovidovye perfectivy); to analyze
constraints on the filling of slots in a construction and how this affects the meaning of
the construction, taking as an example the Genitive construction of shape and the spatial
construction with the preposition poverkh ‘up and over’.

The quantitative corpus-based techniques used for the analysis vary from simple de-
scriptive statistics (e. g., absolute frequencies, percentages, measures of the central ten-
dency and outliers) to exact Fisher test and logistic regression. We claim that the vector
modeling approaches to quantitative grammatical studies in theoretical linguistics are no
less effective than in computational linguistics, where they have become a standard tool.

Olga Lyshevskaya is Professor in the School of Linguistics, Higher School of Eco-
nomics in Moscow, and Senior Researcher in Vinogradov Institute of the Russian Lan-
guage, Russian Academy of Sciences. She is author of Semantics of number in Russian
[Semantika russkogo chisla] (2004), Frequency dictionary of contemporary Russian
based on the Russian National Corpus data [Chastotnyj slovar’ sovremennoho russkogo
jazyka (na materialakh Nacional’nogo korpusa russkogo jazyka)] (with Serge Sharoff,
2009), and Why Russian aspectual prefixes aren’t empty: Prefixes as verb classifiers
(with Laura Janda, Anna Endresen, Julia Kuznetsova, Anastasia Makarova, Tore Nesset,
Svetlana Sokolova, 2013).