CoCoCo: Online Extraction of Russian Multiword Expressions

Mikhail Kopotev[1] **Llorenç Escoter**[2] **Daria Kormacheva**[1]
Matthew Pierce[2] **Lidia Pivovarova**[2] **Roman Yangarber**[2]
[1]University of Helsinki, Department of Modern Languages
[2]University of Helsinki, Department of Computer Science

Abstract

In the CoCoCo project we develop methods to extract multi-word expressions of various kinds—idioms, multi-word lexemes, collocations, and colligations—and to evaluate their linguistic stability in a common, uniform fashion. In this paper we introduce a Web interface, which provides the user with access to these measures, to query Russian-language corpora. Potential users of these tools include language learners, teachers, and linguists.

1 Introduction

We present a system that automatically extracts selectional preferences from a corpus. For a given word, the system finds its selectional preferences, both lexical and grammatical, using algorithms described in (Kopotev et al., 2013; Kormacheva et al., 2014). The system[1] is developed as a part of CoCoCo Project: *Collocations, Colligations, and Corpora*. The system has two important features. First, it allows users to identify selectional preferences, based on a large underlying corpus on-line, in real time, rather than relying on pre-computed lists of multi-word expressions (MWEs). Second, it treats MWEs of various kinds—idioms, multi-word lexemes, collocations and colligations—in a uniform fashion, returning MWEs of all these types in response to a given query.

These features make the system useful for studying a wide variety of linguistic phenomena, depending on the queries formulated by users. For example, in response to a query such as "preposition plus any following word," the system may produce on output a list of nominal *cases* that can be used with (are governed by) the preposition; or a list of most stable phrases with this preposition; or both. The list may contain idioms or collocations that a learner must memorise by rote. The quality of the algorithm is: F-measure 92% for the grammatical preferences task (Kopotev et al., 2013), average precision 24.25% for the lexical preferences though it depends on the queries: for some queries precision is much higher, up to 75% (Kormacheva et al., 2014). An expert linguist may use the system, e.g., to find patterns of use of the so-called "second genitive" case.[2] All queries are processed using the same algorithm; there is no difference between these use-cases in terms of implementation. The system currently works with Russian-language data, but in principle the algorithms and the user interface (UI) can be applied to other typologically similar languages.

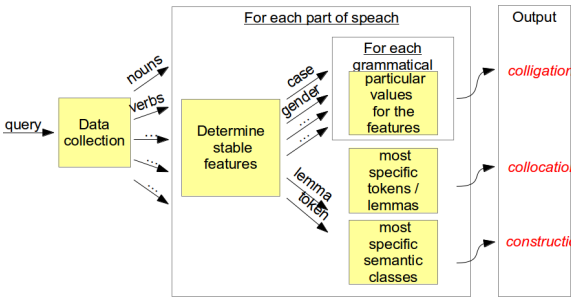

Figure 1: System overview.

From the theoretical perspective, we follow the recent *constructional grammar* approach, where the language is considered as a *constructicon* (Goldberg, 2006), i.e., an inventory of constructions or patterns that predefine both the grammatical and the lexical selectional preferences of words. Distinguishing *collocations*, i.e., "co-occurrences of words" from *colligations*, i.e., "co-occurrence of word forms with grammatical phe-

[1]Accessible at http://corpussearch.cs.helsinki.fi

[2]This case in many instances syncretizes with the normal ("first") genitive, but in many instances does not—it behaves like the partitive case in some languages (e.g., Finnish).

nomena" (Gries and Divjak, 2009) is not always a simple task. There is no clear distinction between various types of word combinations, since they can be simultaneously a collocation and a colligation—this type of MWE is called *collostruction* in (Stefanowitsch and Gries, 2003). Thus our main focus is to find "the underlying cause" for the frequent co-occurrence of certain words: whether it is due to their morphological categories, or to lexical compatibility, or both.

2 Program Overview

The general overview of the system is shown in Figure 1. The system takes as input a query—an N-gram (currently of length 2–4)—where one of the positions is a sought variable, and all positions may have additional, optional grammatical constraints. The constraints may include certain properties, e.g., part of speech (POS), or case, etc. Thus, the query is a pattern. The aim is to find the most stable lexical and grammatical features that match this pattern.

The algorithm finds all words in the corpus that match the pattern, and first groups them according to their POS. Then, for every POS, the system determines the most stable features, which include grammatical categories (case, gender, etc.), tokens, and lemmas. To find the most stable features we exploit the difference between the distribution of the feature values in the pattern vs. distribution in the corpus overall, using a measure based on Kullback-Leibler Divergence (Kopotev et al., 2013).

Having specified the most stable categories, we compute various frequencies to find particular values for these categories (Kormacheva et al., 2014). In this step, grammatical categories are processed separately from tokens and lemmas, since tokens and lemmas have significantly different distributional properties than grammatical categories. The output of the system are colligations and collocations for a given pattern. The combinations of the pattern with the most stable *semantic* classes (constructions) are currently not included in the current version of the on-line tool.

Currently we use two corpora: a (manually) morphologically disambiguated sub-corpus of the Russian National Corpus (Rakhlina, 2005) and the Russian Internet Corpus (Sharoff and Nivre, 2011). The former contains approximately 6 million tokens; from this corpus it is possible to get

Figure 2: On-line interface.

selectional preferences for the most frequent Russian words. The latter corpus contains almost 150 million tokens and is automatically annotated; this corpus may be used to investigate selectional preferences for less frequent words.

3 User Interface

We have implemented a simple graphical interface (GUI) to construct query patterns and obtain results as ordered lists of grammatical and lexical features, Figure 2. Although we show to the user only several most significant results, the algorithm needs to find in the corpus all possible combinations for a given pattern. Since the corpora are large, these would be impossible to manage using plain-text search. Thus, all bi-grams and tri-grams from a corpus are stored in a MySQL database; for the Russian Internet corpus we removed from the data all bi-grams and trigrams that appear in the corpus only once. We use indexing and database optimisation to be able to process user queries on the fly.

The interface has an "Export" function for downloading the complete system output, i.e., the full list of examples matching the pattern in the corpus, ordered according to the measures developed for this task. This output is organized as a set of files in CSV format; these files can be viewed in a spreadsheet, e.g., by users without advanced computational skills. We expect that the export function will be used by professional linguists, while language learners will find that the GUI provides sufficient information for their needs.

Some other functions, such as, for example, batch processing of a set of queries, are currently developed as a command line script and not available for the users outside the CoCoCo team. We

plan to include them into future versions of the interface.

Acknowledgements

The CoCoCo project is partially financed by the BAULT research community (University of Helsinki) and Centre for International Mobility CIMO (Finland). We would like to thank E. Rakhilina, O. Lyashevskaya and S. Sharoff for providing corpora for this project. We thank Ekaterina Nironen for help with Web site design.

References

Adele Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press, USA.

Stefan Th. Gries and Dagmar Divjak. 2009. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. *New Directions in Cognitive Linguistics*, pages 57–75.

Mikhail Kopotev, Lidia Pivovarova, Natalia Kochetkova, and Roman Yangarber. 2013. Automatic detection of stable grammatical features in N-grams. In *9th Workshop on Multiword Expressions (MWE 2013), NAACL HLT 2013*, pages 73–81.

Daria Kormacheva, Lidia Pivovarova, Mihail Kopotev, et al. 2014. Automatic collocation extraction and classification of automatically obtained bigrams. In *Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014)*.

Ekaterina Rakhlina, editor. 2005. *Nacionalnyj korpus russkogo jazyka 2003–2005*.

Serge Sharoff and Joakim Nivre. 2011. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. *Komputernaja lingvistika i intellektualnye tekhnologii: Po materialam Mezhdunarodnoj konferencii Dialog(Bekasovo, 25-29 maja 2011)*, pages 591–604.

Anatol Stefanowitsch and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2):209–243.