ConFarm: Extracting Surface Representations of Verb and Noun Constructions from Dependency Annotated Corpora of Russian

Nikita Mediankin
Institute of Formal and Applied Linguistics
Charles University in Prague, Czech Republic
Faculty of Mathematics and Physics
12800, Praha 2, Sekaninova 14
nikita.medyankin@gmail.com

Abstract

ConFarm is a web service dedicated to extraction of surface representations of verb and noun constructions from dependency annotated corpora of Russian texts. Currently, the extraction of constructions with a specific lemma from SynTagRus and Russian National Corpus is available. The system provides flexible interface that allows users to fine-tune the output. Extracted constructions are grouped by their contents to allow for compact representation, and the groups are visualised as a graph in order to help navigating the extraction results. ConFarm differs from similar existing tools for Russian language in that it offers full constructions, as opposed to extracting separate dependents of search word or working with collocations, and allows users to discover unexpected constructions as opposed to searching for examples of a user-defined construction.

1 Introduction

Certain modern schools of linguistic thought focus on constructions as the means of investigating word meaning. This paradigm, along with rapidly developing capabilities for data-driven research, have recently spawned numerous studies of Russian constructions. For these, specialized resources and tools are required, such as manually annotated frame banks and lexicons, tools for automated or semi-automated expansion of said frame banks, as well as tools for extraction of constructions from large corpora.

The main goal of the presented system is to provide linguists with the means for automatic extraction of verb and noun constructions from dependency annotated treebank of Russian texts. The scope of the system does not include semantic frame labeling, and is restricted to the extraction of surface representation. One of the supposed applications of the system is to help in ongoing development of Russian FrameBank (Lyashevskaya, 2010) by both adding examples to existing constructions and discovering new ones.

2 Difference from Existing Systems

ConFarm differs from similar existing tools that can be used for Russian language, such as SketchEngine (`https://www.sketchengine.co.uk/`), RNC Sketches (`http://ling.go.mail.ru/synt/`), and search in syntactically annotated part of Russian National Corpus (`http://ruscorpora.ru/search-syntax.html`), in the following aspects:

1. For each sentence with search word, ConFarm provides a combination of all extracted dependents. Therefore, it offers full constructions, as opposed to extracting dependents of search word separately or working with collocations.

2. The existing tools mostly allow users to search for examples of a user-defined construction, while ConFarm can be used to discover unexpected constructions by leaving the extraction option about the desirable syntactic relations unspecified in the interface.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 238–242, Osaka, Japan, December 11-17 2016.

3 Corpora

ConFarm allows to extract constructions from two corpora, SynTagRus in its 2015 state (`http://ruscorpora.ru/instruction-syntax.html`), and recent dump of Russian National Corpus (`http://www.ruscorpora.ru`). SynTagRus is a manually annotated dependency treebank of Russian texts. It was automatically converted both for the use by ConFarm and to provide training for MaltParser model used in RU Syntax NLP pipeline (`http://web-corpora.net/wsgi3/ru-syntax/`). Texts from Russian National Corpus were automatically annotated using RU Syntax pipeline. The details on SynTagRus conversion, and RU Syntax pipeline can be found in (Medyankin, Droganova, 2016).

4 Interface

ConFarm extraction page is used both for specifying extraction options and for presenting the results. It allows user to enter lemma, choose part of speech (currently verb or noun), impose a number of restrictions, and choose a number of options for post processing of extracted constructions. Screenshot of the interface is shown in Figure 1. It should be noted that if nothing is specified in 'only with' options, that means no restriction is imposed, e.g., if 'only with syntactic relations' field is left blank, constructions with any syntactic relations will be extracted, thus allowing to discover unexpected constructions.

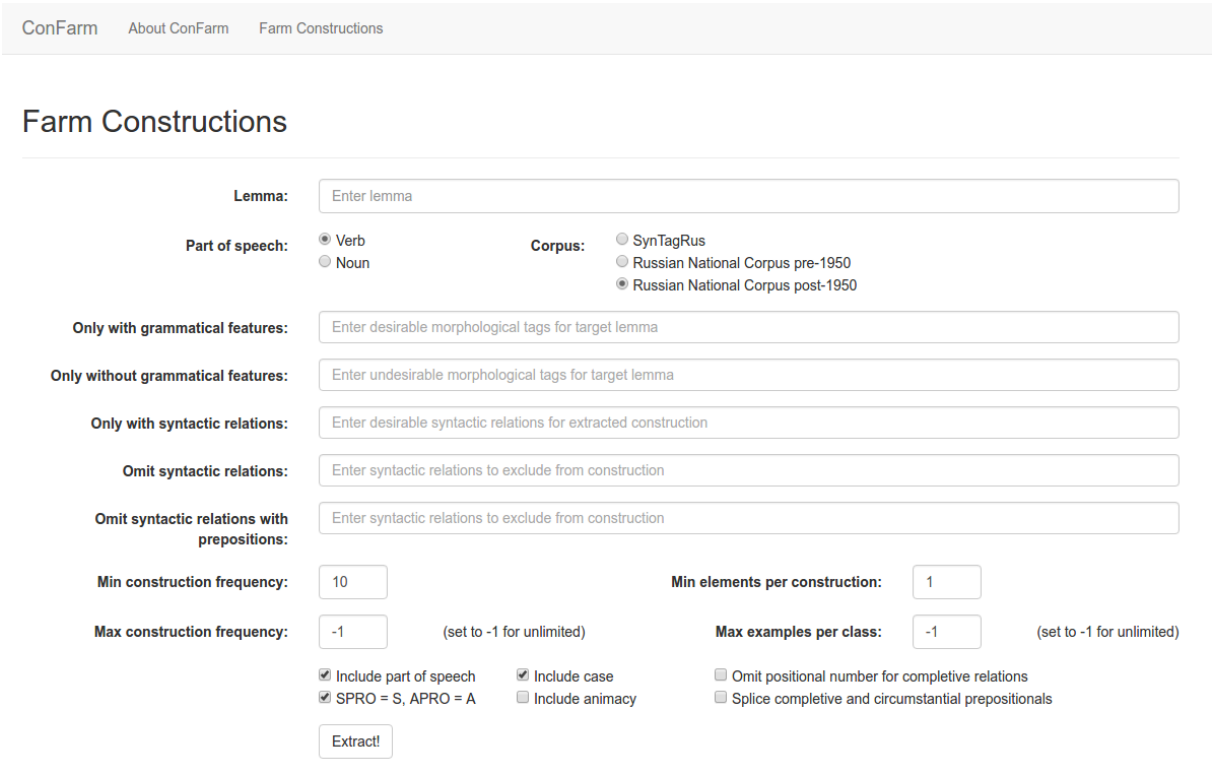

Figure 1: Extraction interface.

The results are presented as both graph and list of extracted constructions grouped by construction contents. Each entry in the list is expandable to show all extracted examples. Each example is shown as a full sentence with extracted construction marked in color. A click on a word opens a popup with information about its lemma, tags, head, and dependency relation label. Figure 2 shows a partial list of constructions extracted for verb *грузить* 'load' from pre-1950 subcorpus of Russian National Corpus.

Figure 2: Partial example of extracted constructions list.

5 Extraction and Classification

Extraction process is rule-based and is performed by a Python3 module specifically written for this purpose.

Immediate and prepositional dependent of search word is always extracted, unless user specifically states in extraction options to exclude dependents with this dependency label. This allows users to fine-tune the balance between recall and precision.

Extraction of other parts of construction is based on a set of rules designed to prevent overextraction and includes additional extraction of dependents of search word's head if it is a verb or a short adjective, and extraction of potential object at the start of coordinated or subordinated chain.

Extracted constructions are grouped by the set of dependency labels present among the parts of the construction. These groups are then viewed as a partially ordered set by inclusion and their relationship is visualized by a Hasse diagram to help navigating the extraction results. The example of the diagram for the constructions with verb *грузить* 'load' extracted from post-1950 subcorpus of Russian National Corpus is shown in Figure 3 (only constructions with frequency more than 10 in the corpus were considered).

Figure 3: An example of classification diagram.

6 Evaluation and Discussion

For the purposes of evaluation, the following test has been conducted. 200 examples of different verb constructions with their verb in any form but participle, each with an illustrative chunk of text from Russian National Corpus, were chosen at random from Russian FrameBank (`http://framebank.ru/`). Only arguments were considered part of construction, no adjuncts were included. Each illustrative chunk was then ran through the same stages as though it was annotated for ConFarm and a construction was extracted from it, i.e., annotated with RU Syntax and passed to the Python3 function used to extracts constructions from a sentence. The following settings were chosen as a tradeoff between precision and recall: exclude circumstantial dependents without preposition, exclude parenthetic, delimitative, and expository dependents. This was done in order to reduce adjuncts in the results. Since no exactly similar systems are available for Russian language to compare the results to, a simple baseline was developed: extract

all nouns, infinitive verbs, and prepositions directly preceding them within ±5 context window or sentence boundaries, whatever is met first.

FrameBank	**Baseline**	**ConFarm**
subject	Noun nom	Noun nom, predicative
	Verb inf	Verb inf, predicative
object	Noun acc	Noun acc, completive
	Verb inf	Verb inf, completive
periphery	Noun other case	Noun other case, completive
	Prep + Noun other case	Prep + Noun other case, completive

Table 1: FrameBank to Baseline to ConFarm match for labeled scores.

The results were then manually compared with FrameBank annotations. First, unlabeled scores were calculated: (1) if given token is present both in FrameBank annotation and extracted construction, it is considered true positive, disregarding its dependency label and FrameBank labeling; (2) if it is present in FrameBank annotation, but not in extracted construction, it is considered false negative; (3) if it is not present in FrameBank annotation, but is present in extracted construction, it is considered false positive; (4) if it is not present in FrameBank annotation, nor in extracted construction, it is considered true negative. Unlabeled precision, recall and accuracy were then calculated following standard definitions. Second, labeled scores were calculated: same as above but given token was only considered a hit if (a) its case (for nouns) or infinitiveness (for verbs) matched FrameBank, and (b) its dependency label corresponded to its FrameBank rank as shown in Table 1. For baseline, only (a) was considered when calculating labeled scores. The scores are shown in Table 2.

	Unlabeled			Labeled		
	precision	recall	accuracy	precision	recall	accuracy
Baseline	51%	77%	85%	44%	67%	82%
ConFarm	75%	79%	93%	64%	68%	89%

Table 2: Test results.

With both labeled and unlabeled scores, ConFarm showed much higher precision and slightly higher recall, compared to the baseline. Detailed examination of the results showed that better precision was due to ConFarm filtering out irrelevant nouns and infinitives, and the recall was higher because of detected distant parts of construction that did not get to the context window, but not by the large margin because a number of relevant completive dependents were erroneously marked as circumstantial and therefore filtered out.

7 Availability

ConFarm web-service is available for unconditional use at `http://www.confarm.online`.

8 Conclusion

This article presented a web-service ConFarm that provides extraction and initial classification of surface representations of verb and noun constructions from two dependency annotated Russian corpora: SynTagRus and Russian National Corpus, the latter of which was automatically dependency annotated specifically for the purpose of using it in ConFarm. The web-interface allows users to fine-tune the output by specifying a number of various extraction options. The system was evaluated on 200 different verb constructions from Russian FrameBank and results compared to a simple baseline set up without using dependency annotation. For both labeled and unlabeled

scores, ConFarm showed much higher precision and slightly higher recall than the baseline. Further improvements can be made to the system by both obtaining better automated annotation for Russian National Corpus and by refining the rules for extracting parts of the construction that are not immediate or prepositional dependents of the search word.

Acknowledgements

This work was partially funded by the Ministry of Education, Youth and Sports of the Czech Republic under the project SVV project 260 333. It used language resources stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

Olga Lyashevskaya. 2010. *Bank of Russian constructions and valencies.* LREC 2010. Malta, Valletta, May 19-21, 2010.

Nikita Medyankin, Kira Droganova. 2016. *Building NLP Pipeline for Russian with a Handful of Linguistic Knowledge.* Online proceedings of the Workshop "Computational linguistics and language science" (CLLS) Moscow 2016, CEUR Workshop Proceedings (in print, manuscript is available at `http://web-corpora.net/wsgi3/ru-syntax/static/downloads/Medyankin_Droganova_CLLS_2016.pdf`).