Crowdsourcing for the Russian Morphological Lexicon

Vladimír Benko[a] and Victor Zakharov[b]

[a] *JULS, Slovak Academy of Sciences, Panská 26, Bratislava, 811 01, Slovakia*
[b] *St Petersburg University, Universitetskaya emb. 7-9, Saint Petersburg, 199034, Russia*

Abstract

We present an on-going experiment aimed at improving the results of Russian PoS tagging by means of increasing the size of morphological lexicon that is used for training the respective tagger(s). The frequency list of out-of-vocabulary (OOV) word forms along with the tags and lemmas assigned by the guesser is manually checked, corrected and classified by students in the framework of assignments, so that valid lexical items candidates for inclusion into the morphological lexicon could be identified. We expect to improve the lexicon coverage by the most frequent proper names and foreign words, as well as to create an auxiliary lexicon containing the most frequent typos.

Keywords

crowdsourcing, Russian POS-tagging, out-of-vocabulary words

1. Introduction

Assuming that one of the main features of a representative text corpus is its size, then a 100-million token corpus, considered a standard at the beginning of the century, now often appears to be insufficient to collect relevant statistical data. As soon as the need for larger corpora has been recognized, it became clear that the requirements of the linguistic community cannot be fully satisfied by the traditional methods of building corpora. At the turn of the new millennium, the idea of Web as Corpus (WaC), i.e., creation of language corpora based on the web-crawled data has been born, for the first time explicitly articulated by Adam Kilgarriff [1, 2]. In early 2000s, a community called WaCky! was established by a group of linguists and IT specialists to develop tools for creation of large-scale web-crawled corpora. During the period of 2006–2009, several WaC corpora were created and published [3]. Since then, several other initiatives emerged [4-8], with one of them also being the Aranea Web Corpora Project [9].

The Aranea family presently consists of (comparable) web corpora created for more than two dozens of languages and language varieties. The corpora bear Latin names denoting the Language and size, with two sizes being typically available. The Maius ("larger") series corpora contain 1.25 billion tokens, i.e., approximately 1 billion words (tokens starting with an alphabetic character). Each Minus ("smaller") corpus represents a 10% random sample of the respective Maius corpus. For languages spoken in more than one country, corpora for region-specific language varieties may exist. For Russian, for example, the Araneum Russicum consists of Russian texts downloaded from any internet domain, Araneum Russicum Russicum contain only texts extracted from the .ru and .pф .su domains, and Araneum Russicum Externum are based on texts from "non-Russian" domains, such as .ua, .by, .kz, etc. For more details about the Aranea Project see [9, 10]. For some languages, a Maximum class corpora are also created applying the strategy "as much as can get". The largest corpus withn the Aranea family is Russicum Maximum containing almost 20 billion tokens.

To create a web corpus, we usually have to perform (in a certain sequence) operations as follows:

- Downloading large amounts of data from the Internet, extracting the textual information, normalizing encoding

1IMS 2021 - International Conference "Internet and Modern Society", June 24-26, 2021, St. Petersburg, Russia
EMAIL: vladimir.benko@juls.savba.sk (A. 1); v.zakharov@spbu.ru (A. 2);
ORCID: 0000-0002-6733-8386 (A. 1); 0000-0003-0522-7469 (A. 2)

- Identification the language of the downloaded texts, removing the "incorrect" documents
- Segmenting the text into paragraphs and sentences
- Removing duplicate contents (identical or partially identical text segments)
- Tokenization—segmenting the text into words
- Linguistic (morphological, and possibly also syntactic) annotation—lemmatization and PoS tagging
- Uploading the resulting corpus into the corpus manager (i.e., generating the respective index structures) that will make the corpus accessible for the users.

Our paper is devoted to the morphological tagging of input texts and, narrower, to the processing of out-of-vocabulary (OOV) tokens.

2. Morphosyntactic annotation

From the very beginning of the Aranea Project, only tools with an open-source or free license have been used for all processing. As there are many languages to be processed, for morphosyntactic annotation tools with many language models were preferred. This was especially the case of TreeTagger, that was easy to integrate into our processing pipeline [11, 12].

Despite being a rather old tool, TreeTagger is still being used by many projects. Its main advantage, from our perspective, is the processing speed than can be even by the order of magnitude faster than other tools for the same language. There are, however, some disadvantages as well. The quality of language models provided by its author varies from one language to another, depending on the training data and morphological lexicon available. Perhaps the greatest deficiency of TreeTagger is absence of any procedure that would guess lemmas for out-of-vocabulary (OOV) lexical items – those are simply tagged as "unknown", leaving decision of further processing to the user.

Several Russian language models for TreeTaggers are available, for our work we rely on that provided by Serge Sharoff. Though its coverage, compared to other languages, is fairly high, the morphological lexicon based on Zaliznyak dictionary and manually disambiguated subcorpus of the Russian National Corpus naturally cannot cover all lexical items appearing in a "fresh" web-crawled corpus.

Russian belongs to languages with several taggers available, so an idea of looking for an alternative is quite straightforward. None of them, however, can be simply declared as "better" – each of them has some drawbacks as well. In our experiments we were using UDPipe [13], a tool developed in the framework of the Universal Dependencies Project [14]. It is able to guess lemmas but, unlike TreeTagger, the UDPipe does not use morphological lexicon at all and all lemmas are guessed, even for lexical items present in the training data. Out of the language models available, we opted for that trained on the SynTagRus treebank [15].

The third tool included in our work was CSTlemma [16, 17], a high performance lemmatizer with a language model provided by its author.

In an attempt to improve the results of annotation, we are planning to apply steps as follows.

1 Use the "ensemble tagging" approach, i.e., annotate the corpus by several different tools.
2 Aggregate the results by means of manually written rules.
3 Manually disambiguate the annotations for most frequent OOV lexical items
4 Use the disambiguated list to amend the morphological lexicon for next step of annotation.

The current paper describes the very first phase of or experiment targeted on correcting the lemma and PoS tags by means of crowdsourcing.

3. Crowdsourcing

“Crowdsourcing" is a relatively recent concept that encompasses many practices. This diversity leads to the blurring of the limits of crowdsourcing that may be identified virtually with any type of Internet-based collaborative activity, such as co-creation or user innovation [18]. In their paper, authors define eight characteristics typical for crowdsourcing as follows:

- There is a clearly defined crowd (a)
- There exists a task with a clear goal (b)
- The recompense received by the crowd is clear (c)
- The crowdsourcer is clearly identified (d)
- The compensation to be received by the crowdsourcer is clearly defined (e)
- It is an online assigned process ofparticipative type f)
- It uses an open call of variable extent (g)
- It uses the Internet (h)

From this perspective, language data annotation performed by students in the framework of the end-of-term assignments can well be considered “crowdsourcing”, even if only some of the above characteristics apply. It is also worth noting that, according to our experience, students appreciate the feeling that their work may be useful not only as a tool for classification.

4. The Task

The OOV lexical items observed in our corpora are of different nature. Besides the “true neologisms”, i.e., words qualifying for inclusion even into the traditional dictionary, proper nouns (such as personal and geographical names) and their derivates, we can find also items traditionally not considered as “words” – various abbreviations, acronyms and symbols, URLs or e-mail addresses, parts of foreign language quotations and – above all – all sorts of “typos” and “errors”. Inflected word forms apply to almost all previously mentioned categories, which makes the whole picture even more complex.

In the following text we present an experiment aimed at amending the morphological lexicon used for training the language model(s) by a manually validated list of most frequent OOV items derived from an annotated web corpus. The annotation is to be performed by graduate students of the Mathematical Linguistics Departmernt of the Saint-Petersburg University in the framework of end-of-term assignment for the “Corpora in NLP” subject.

Having only limited “human power” (14 students in total) at hand, we decided to follow the three-fold setup (i.e., each item to be annotated by three independent annotators) and make the task as simple as possible. This is why the annotators were not expected to check all the morphological categories provided by the respective tags, and they were asked to decide only on two parameters - lemma and word class (part of speech).

5. The Data

In the first step, we used data from and the Aranea TreeTagger pipelines, and subsequently merged into a single vertical file. Then, we converted the original MTE morphological tags to “PoS- only” tags and produced a frequency list of all lexical items indicated as OOV by both taggers. After deleting the unused parameters, the resulting lists contained the frequency, word form, lemma assigned by the CSTLemma and UDPipe taggers and PoS information derived from the tag assigned by TreeTagger (aTag, using the AUT notation). This decision has been motivated by an observation that TreeTagger is typically more successful in assigning morphological categories for unknown words than others.

As we naturally could expect to be able to process only the rather small part of the list, after some experimenting with various thresholds, we decided to pass into annotation only the most frequent items. This meant that each annotator would process approximately 1000 items.

The example of the source data (in alphabetical order, after applying the frequency cut-off) is shown in Table 1.

Table 1
Source Data

Freq	Word	aTag (TreeTagger)	uLemma (UDPipe)	uAtag (UDPipe)	CLemma (CSTLemma)
326	Росстата	Nn	Росстат	Nn	Росстат
116	Ростех	Nn	Ростех	Dt	Ростеха
182	Ростехнадзора	Nn	Ростехнадзор	Nn	Ростехнадзор
117	Ростова-на-Дону	Nn	Ростова-на-Дон	Nn	ростова-на-дон
202	Ростове-на-Дону	Nn	Ростове-на-Дон	Nn	ростове-на-дон
107	Ростове-на-Дону	Nn	ростов-на-дону	Nn	ростове-на-дон
156	Ростов-на-Дону	Nn	Ростов-на-Дон	Nn	ростов-на-дон
202	ротовую	Aj	ротовый	Aj	ротовый
105	роуминг	Nn	роуминг	Nn	роуминг
83	роуминга	Nn	роуминг	Nn	роуминг
176	роутер	Nn	роутер	Nn	роутер
104	РПЛ	Zz	РПЛ	Nn	рплый
227	РСА	Nn	РСА	Nn	РС
287	РСО-Алания	Nn	РСО-Алания	Nn	РСО-Алания
114	рубцов	Nn	рубец	Nn	рубец
220	руд	Nn	руда	Nn	руд
95	руду	Nn	руда	Nn	руда
91	рулонных	Aj	рулонный	Aj	рулонный
99	румяной	Aj	румяный	Aj	румяный
87	РУСАДА	Nn	РУСАД	Nn	РУСАДА
145	РусГидро	Nn	РусГидро	Nn	русгидро
98	Руссель	Nn	Руссель	Nn	Руссель
83	ручках	Nn	ручка	Nn	ручка
212	ручном	Aj	ручной	Aj	ручный

We can observe several phenomena here. While most PoS categories are classified correctly, abbreviation are mostly tagged as “nouns”, but also as “determiners”, or even “punctuation”, and lemma form as well as its capitalization is sometimes guessed correctly, while sometimes not. The result of simple aggregation of the same data can be seen in Table 2.

The overall task for the annotators was to produce correct data for all lines in the table. To minimize the number of necessary keystrokes and to keep track of the changes, the data have been further modified to contain two newly added columns – Lemmb used as a template for correcting the value for Lemma (it is expected that most modifications will occur at the end of the respective string only) and bTag (to be filled only in case of wrong PoS assignment).

As has been already mentioned, each item (line of the table) has to be annotated by three independent annotators. We decided, however, not to split the data in a straightforward way, but to assign each alphabetical segment of the data to three annotators using a rule as follows: each group of four lines will be split into four tuples containing three lines with one missing line form the original group. Moreover, the whole lot of data has been split to three parts, so that each annotator could get three different sections of the alphabet in his or her data.

Table 2
Aggregated annotations, frequency counts discarded, a unique Id added.

Id	Word	Lemma	aTag
ru_003798	Росстата	Росстат	Nn
ru_003799	Ростех	Ростех|Ростеха	Nn|Dt
ru_003800	Ростехнадзора	Ростехнадзор	Nn
ru_003801	Ростова-на-Дону	Ростова-на-Дон|ростова-на-дон	Nn
ru_003802	Ростове-на-Дону	Ростове-на-Дон|ростове-на-дон	Nn
ru_003803	Ростове-на-Дону	ростов-на-дону|ростове-на-дон	Nn
ru_003804	Ростов-на-Дону	Ростов-на-Дон|ростов-на-дон	Nn
ru_003805	ротовую	ротовый	Aj
ru_003806	роуминг	роуминг	Nn
ru_003807	роуминга	роуминг	Nn
ru_003808	роутер	роутер	Nn
ru_003808	РПЛ	РПЛ|рплый	Zz|Nn
ru_003810	РСА	РСА|РС	Nn
ru_003811	РСО-Алания	РСО-Алания	Nn
ru_003812	рубцов	рубец	Nn
ru_003813	руд	руда|руд	Nn
ru_003814	руду	руда	Nn
ru_003815	рулонных	рулонный	Aj
ru_003816	румяной	румяный	Aj
ru_003817	РУСАДА	РУСАД|РУСАДА	Nn
ru_003818	РусГидро	РусГидро|русгидро	Nn
ru_003819	Руссель	Руссель	Nn
ru_003820	ручках	ручка	Nn
ru_003821	ручном	ручной|ручный	Aj

By applying this fairly “sophisticated” assignment scheme, we expected to improve the overall uniformity and quality of the output, as well as to prevent “collaboration” among students, as no two assigned lots were identical.

An excerpt of the data from Table 3 assigned to a single annotator is shown in Table 3.

Table 3
Data to Annotate

Id	Word	Lemma	Lemmb	bTag	aTag
ru_003797	Росстат	Росстат	Росстат		Nn
ru_003799	Ростех	Ростех|Ростеха	Ростех|Ростеха		Nn|Dt
ru_003800	Ростехнадзора	Ростехнадзор	Ростехнадзор		Nn
ru_003801	Ростова-на-Дону	Ростова-на-Дон|ростова-на-дон	Ростова-на-Дон|ростова-на-дон		Nn

Id	Word	Lemma	Lemmb	bTag	aTag
ru_003803	Ростове-на-Дону	ростов-на-дону|ростове-на-дон	ростов-на-дону|ростове-на-дон		Nn
ru_003804	Ростов-на-Дону	Ростов-на-Дон|ростов-на-дон	Ростов-на-Дон|ростов-на-дон		Nn
ru_003805	ротовую	ротовый	ротовый		Aj
ru_003807	роуминга	роуминг	роуминг		Nn
ru_003808	роутер	роутер	роутер		Nn
ru_003809	РПЛ	РПЛ|рплый	РПЛ|рплый		Zz|Nn
ru_003811	РСО-Алания	РСО-Алания	РСО-Алания		Nn
ru_003812	рубцов	рубец	рубец		Nn
ru_003813	руд	руда|руд	руда|руд		Nn
ru_003815	рулонных	рулонный	рулонный		Aj
ru_003816	румяной	румяный	румяный		Aj
ru_003817	РУСАДА	РУСАД|РУСАДА	РУСАД|РУСАДА		Nn
ru_003819	Руссель	Руссель	Руссель		Nn
ru_003820	ручках	ручка	ручка		Nn
ru_003821	ручном	ручной|ручный	ручной|ручный		Aj

Note that the "missing" every third Id results from the assignment scheme.

6. The Crowd Annotation

The split data has been uploaded as excel spreadsheets to a shared Google disk and assigned randomly to the respective annotators. The task has been assigned in the middle of the semester, after the students already got acquainted with the basic concepts of corpus morphosyntactic annotation and acquired the elementary querying skills. The instructions for annotating the data as they are presented in Table 3 were as follows.

A Only Lemmb and bTag columns may be modified.
B If both Lemma and aTag values are correct, nothing has to be done.
C If aTag value is wrong, the correct value should be inserted in bTag.
D If Lemma value is wrong, it should be corrected in Lemmb.
E If the word form is obvious typo (missing or superfluous letter, exchanged letters), or the word does not contain the necessary diacritics, the correct lemma marked by an asterisk should entered in Lemmb.
F If the correct word form cannot be reconstructed by simple editing operations, i.e., cannot be recognized (e.g., part of the word as a result of hyphenation), the value of bTag will be "Er" (error).
G If the word form is obvious foreign word, the value of bTag will be "Yx".
H It is not necessary to evaluate whether the word form is "literary" - words of "lower" registers (such as slang) also have "correct" lemmas.

The annotators were also instructed to check all "non-obvious" items by querying the corpus and analyzing the respective contexts. The initial training was performed during one teaching lesson in a computer lab, so that possibly all frequent problems could be explained.

7. Linguistic aspects of Russian tagging

Obviously, recommendations for reannotation of OOV word forms should be not so much technical as linguistic. They should analyze not only typical obvious cases, but also problems. In this case, we should proceed from the following:

- - Russian grammar rules
- - considering tokenization rules when annotating a corpus
- - contexts of using this token in the corpus
- - frequency data on certain uses.

The development and description of such an instruction is a matter of the future and a topic for a separate article, here we give some problems that cause difficulties during annotation.

1. Foreign words should be processed depending on the context: if they appear as part of a foreign language expression (quotation), then the correct PoS tag is "Yx". However, if they are part of a Russian-language phrase, for example, in the meaning of a noun (usually these are proper nouns), then it is reasonable to mark them with the tag "Nn" (or other relevant part of speech).
2. Many languages and taggers have problems with abbreviations. We can say that they are difficult for grammar as such. Usually, abbreviations include words written in capital letters, however, there are many other options. For example. The abbreviation ВУЗ (high school) is widely spelled in small letters (вуз), which has actually become a noun. There are a large number of standardized and non-standardized abbreviations such as *д-р, (doctor)*, *изд-во (*Publishing House) etc. Abbreviations like МХАТ, *НИИМаш* are often inflected as nouns and in fact, they are, without losing the spelling in large letters. Some abbreviature have several standard lemmas (spellings), eg, КзоТ, *КЗОТ. (L*abor Code, Labor Code).
3. It is desirable to include proper names in the morphological dictionary. The question arises, all or quite often used? So, if the dictionary contains the masculine name Давид , should the feminine name Давида be included with a different declension paradigm? Almost any adjective or common noun can occur as a surname. Should they be presented in the morphological dictionary as separate lexemes? Those common nouns in indirect cases, apparently, need to match two lemmas, for example, the word form Котов (surname) will receive a tag of a noun and two lemmas Котов|кот, or Юнг lemmas Юнг|юнга, the word form in genetivs case Серебряной (surname) gets lemmas Серебряная|серебряный and tags Nn | Aj.
4. The participles in MTE are carried out to the verb lemma (выделенный — выделить), but there are many cases when, along with the verbal lemma, an adjective lemma must also be indicated: *добавленная стоимость* - добавить|*добавленный Vb|Aj*.

8. First Results and Problems

The source data consisted of 5,040 producing 15,120 lines to annotate by 15 students. I.e., each of them had to process 1,008 lines. As only 14 files have been returned, the missing file has been reassigned to a student from a different group.

The resulting data has not been processed completely yet, but the first analysis looks promising – see table 4 and 5.

Table 4
Results of Annotation

	Count	%
Source lines	5,040	100.00 %
Triple Agreement both on Lemma and PoS	4,202	83.37%
Double Agreement both on Lemma and PoS	649	12.88%
Triple Agreement on Lemma	4,448	88.25%
Double Agreement on Lemma	498	9.88%

Table 5
Annotated Data PoS Distribution

PoS	Count	%
Nn	20043	73.86
Aj	5174	19.07
Pn	46	0.17
Nm	27	0.10
Vb	464	1.71
Av	261	0.96
Pp	8	0.03
Cj	10	0.04
Ij	42	0.15
Pt	24	0.09
Ab	185	0.68
Xy	1	0.00
Yx	490	1.81
Er	343	1.26
?	17	0.06
	27135	100.00

9. Conclusions and Further Work

There were several goals to be achieved by the annotation. Firstly, we would like to produce a validated list of most frequent neologisms to be included in the morphological lexicon; in this stage, we even do not expect to generate full paradigms for those lexical items. Secondly, we wanted to get the list of the most frequent typos and other types of errors that could also be used as a supplement to that lexicon, but also as source data for a future system for data normalization. And lastly, we also wanted to obtain a list of most frequent foreign lexical items appearing in Russian corpus data.

Although the detailed analysis of the annotated data is yet to be performed, some conclusions can be seen already. They can be summarized as follows:

(1) The Annotation Guidelines must be as precise as possible, showing not only the typical problems and their solutions, but also the seemingly "easy" cases. One-page instruction, as it was in our case, is definitely not sufficient.

(2) The most common errors were associated with the treatment of proper nouns. An automatic procedure based on frequencies of lower/uppercased word forms would most likely perform better.

(3) The other common issue was the proper form of lemma for adjectives (it should be masculine and nominative singular). As the morphology of the Russian adjectives is fairly regular, a procedure to fix it automatically would be feasible.

(4) One of the fairy frequent PoS ambiguity in our data was the "Nn"/"Yx" (noun/foreign) case. The manually annotated data, however, show that the real number of "foreign" is rather low, yet it introduces a lot of noise into the annotation process. It would therefore be reasonable to substitute all tags for "foreign" with that of "nouns" in the future annotation.

In the near future, besides the new round of a similar annotation effort with an improved setup, we would like to combine its results with those obtained in the framework of the ensemble tagging experiment.

10.Acknowledgment

Authors wish to express their sincere gratitude to the 1st year Master program students of the SPbU Mathematical Linguistics Department for their valuable help in annotating the data.

11.References

[1] A. Kilgarriff, Web as corpus, *Proc. of Corpus Linguistics 2001 conference*, Lancaster University. Lancaster: UCREL, 2001, pp. 342–344.

[2] A. Kilgarriff and G. Grefenstette, Introduction to the Special Issue on Web as Corpus, Computational Linguistics, vol. 29, no. 3, pp. 333–347, 2003.

[3] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta, The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation, vol. 43, no. 3, pp. 209–226, 2009.

[4] N. Ljubešić and T. Erjavec, hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene, Text, Speech and Dialogue — 2011. Lecture Notes in Computer Science, Springer, 2011, pp. 395–402.

[5] N. Ljubešić and F. Klubička, {bs,hr,sr}WaC — Web corpora of Bosnian, Croatian and Serbian, Proceedings of the 9th Web as Corpus Workshop (WaC-9), Gothenburg: Association for Computational Linguistics, 2014, pp. 29–35.

[6] M. Jakubíček, A. Kilgarriff, V. Kovář et al., The TenTen Corpus Family, Proceedings of the 7th International Corpus Linguistics Conference, Lancaster: UCREK, 2013, pp. 125–127.

[7] V. I. Belikov, V. P. Selegey, and S. A. Sharov, Prolegomena to the project of the General Internet Corpus of the Russian Language (GIKRYA) [Prolegomeny k proyektu General'nogo internet-korpusa russkogo yazyka (GIKRYA)], Computational Linguistics and Intellectual Technologies: Based on the materials of the annual international conference "Dialogue" (Bekasovo, May 30 - June 3, 2012), issue 11 (18).,Moscow: RGGU Publishing House, 2012, vol. 1, pp. 37–49.

[8] R. Schäfer, F. Bildhauer, Building Large Corpora from the Web Using a New Efficient Tool Chain, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), Istanbul: European Language Resources Association, 2012, pp. 486–493.

[9] V. Benko, Aranea: Yet Another Family of (Compara¬ble) Web Corpora, P. Sojka, A. Horák, I. Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings, LNCS 8655, Springer International Publishing Swit¬zerland, 2014, pp. 257-264.

[10] V. Benko, Two Years of Aranea: Increasing Counts and Tuning the Pipeline, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016), Portorož : European Lan¬guage Resources Association (ELRA), 2016, pp. 4245-4248.

[11] H Schmid, Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.

[12] H Schmid, Improvements in Part-of-Speech Tagging with an Application to German, Proceedings of the ACL SIGDAT-Workshop, Dublin. 1995.

[13] M. Straka, UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task, Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium : Association for Computational Linguistics, 2018, pp. 197–207. DOI: 10.18653/v1/K18-2020. URL: https://www.aclweb.org/anthology/K18-2020.

[14] Universal Dependencies URL: https://universaldependencies.org/

[15] UD Russian SynTagRus 2021 URL: https://universaldependencies.org/treebanks/ru_syntagrus/index.html

[16] B. Jongejan and H. Dorte, The CST Lemmatiser. Center for Sprogteknologi, University of Copenhagen version 2.7, 2005. URL: http://cst.dk/online/lemmatiser/cstlemma.pdf

[17] B. Jongejan and C. Navarretta, CLARIN-DK presents the CST Lemmatizer, 2019. URL: https://www.clarin.eu/blog/clarin-dk-presents-cst-lemmatizer

[18] E. Estellés-Arolas and F. González-Ladrón-de-Guevara, Towards an Integrated Crowdsourcing Definition, Journal of Information Science, 2012, vol. 38, no. 2), pp. 189¬200, doi:10.1177/0165551512437638.