A preliminary constraint grammar for Russian

Francis M. Tyers
HSL-fakultehta,
UiT Norgga árktalaš universitehta,
N-9018 Romsa
francis.tyers@uit.no

Robert Reynolds
HSL-fakultehta,
UiT Norgga árktalaš universitehta,
N-9018 Romsa
robert.reynolds@uit.no

Abstract

This paper presents preliminary work on a constraint grammar based disambiguator for Russian. Russian is a Slavic language with a high degree of both in-category and out-category homonymy in the inflectional system. The pipeline consists of a finite-state morphological analyser and constraint grammar. The constraint grammar is tuned to be high recall (over 0.99) at the expense of low precision.

1 Introduction

This paper presents a preliminary constraint grammar for Russian. The main objective of the constraint grammar is to produce a high recall grammar to serve as input into other natural language processing tasks. There are two reasons to maintain high recall. First, one of the primary applications for this constraint grammar is computer-assisted language learning. In the domain, erroneous analyses can lead to significant frustration for learners. Therefore, it is important to limit disambiguation to cases that can be resolved with high confidence. Second, it is frequently the case that competing readings can be distinguished only by considering idiosyncratic collocational information. For such cases, we expect that probabilistic approaches are both more effective and simpler to implement.

The paper is laid out as follows: section 2 presents a review of the literature on Russian language processing; section 3 gives an overview of ambiguity in Russian; section 4 describes our analysis pipeline; section 5 gives an account of our development process; section 6 presents an evaluation of the system, and sections 7 and 8 present future work and conclusions.

2 Review of literature

State-of-the-art morphological analysis in Russian is primarily based on finite-state technology (Nozhov, 2003; Segalovich, 2003).[1] Almost without exception, all large-scale morphological transducers of Russian are based on the forward-looking *Grammatical Dictionary of Russian* (Zaliznjak, 1977). This dictionary gives fine-grained morphological specifications for more than 100 000 words, including inflectional endings, morphophonemic alternations, stress patterns, exceptions, and idiosyncratic collocations. We developed a morphological transducer based on Zaliznjak's dictionary.[2] This finite-state transducer (FST) generates all possible morphosyntactic readings of each wordform, regardless of its frequency or probability. Because Russian is a relatively highly inflected language, broad coverage is important, but widespread homonymy leads to the generation of many spurious readings, as discussed in Section 3 below. Because of this, one of the foundational steps in Russian natural language processing is homograph disambiguation.

3 Ambiguity in Russian

We identify three different types of morphosyntactic ambiguity: intraparadigmatic, morphosyntactically incongruent, and morphosyntactically congruent. The following examples make use of word stress ambiguity to illustrate each kind of ambiguity.[3] *Intraparadigmatic* ambiguity refers to homo-

[1] Machine-learning approaches have also been successfully applied to Russian, most notably by Sharoff et al. (2008).

[2] Our transducer is implemented using a two-level morphology (Koskenniemi, 1984), and can be compiled using either `xfst` (Beesley and Karttunen, 2003) or `hfst` (Linden et al., 2011)

[3] Written standard Russian does not typically indicate stress position, but knowing stress position is essential for pronunciation. A recent study by Reynolds and Tyers (2015) found that about 7.5% of morphosyntactic ambiguity in a cor-

graphic wordforms belonging to the same lexeme, as shown in (1).

(1) Intraparadigmatic homographs
  a. те́ла *téla* ‘body.$_{\text{SG-GEN}}$’
  b. тела́ *telá* ‘body.$_{\text{PL-NOM}}$’

The remaining two types of ambiguity occur between lexemes. *Morphosyntactically incongruent* ambiguity occurs between homographs that belong to separate lexemes, and whose morphosyntactic values are different, as shown in (2).

(2) Morphosyntactically incongruent homographs
  a. на́шей *nášej* ‘our.$_{\text{F-SG-GEN/DAT/LOC...}}$’
   наше́й *našéj* ‘sew on.$_{\text{IMP-2SG}}$’
  b. доро́га *doróga* ‘road.$_{\text{N-F-SG-NOM}}$’
   дорога́ *dorogá* ‘dear.$_{\text{ADJ-F-SG-PRED}}$’

Morphosyntactically congruent ambiguity occurs between homographs that belong to separate lexemes, and whose morphosyntactic values are identical, as shown in (3).

(3) Morphosyntactically congruent homographs
  a. за́мок *zámok* ‘castle.$_{\text{SG-NOM}}$’
   замо́к *zamók* ‘lock.$_{\text{SG-NOM}}$’
  b. за́мков *zámkov* ‘castle.$_{\text{PL-GEN}}$’
   замко́в *zamkóv* ‘lock.$_{\text{PL-GEN}}$’
   etc.

Table 1 shows the prevalence of each kind of ambiguity. The first column shows the proportion of all tokens in a corpus that have each kind of ambiguity. The second column shows what proportion of ambiguous tokens exhibit each kind of ambiguity. Note that these proportions do not sum to 100%, since a given token may exhibit more than one kind of ambiguity. For example, the wordform *zamkov* has the readings given in (4).

(4) a. замок^1+N+Msc+Inan+Pl+Gen
  b. замок^2+N+Msc+Inan+Pl+Gen
  c. замковый+A+Msc+Sg+Pred

The ambiguity between (4-a) and (4-b) is morphosyntactically congruent, and the ambiguity between (4-a)/(4-b) and (4-c) is morphosyntactically incongruent, so this wordform would be counted for both categories in Table 1.

pus of Russian resulted in stress position ambiguity.

Table 1 shows that most morphosyntactic ambiguity in unrestricted Russian text is rooted in intraparadigmatic and morphosyntactically incongruent ambiguity. Detailed part-of-speech tagging with morphosyntactic analysis can help disambiguate these forms. On the other hand, morphosyntactically congruent ambiguity represents only a very small percentage of ambiguous wordforms, and instead of detailed part-of-speech tagging, it can be resolved by means of word sense disambiguation. Because of this difference, we leave morphosyntactically congruent ambiguity to future work.

Type	**all tokens**	**ambig. tokens**
Intraparadigm.	59.0%	90.9%
Incongruent	27.7%	42.7%
Congruent	1.2%	1.8%

Table 1: Frequency of different types of morphosyntactic ambiguity in unrestricted text

4 Pipeline

4.1 Morphological analyser

The morphological transducer used in this study is primarily based on Zaliznjak's *Grammatical dictionary of Russian*, including the 2001 version's appendix of proper nouns. It also includes neologisms from Grishina and Lyashevskaya's *Grammatical dictionary of new Russian words*, which is intended to be a supplement to Zaliznjak's dictionary with words found in the Russian National Corpus.[4] Example (5) gives some examples of the FST's output.

(5) a. `новый<adj><m><nn><sg><nom>`
   ‘new’
  b. `автомат<n><m><nn><sg><nom>`
   ‘automaton, sub-machine gun’

4.2 Disambiguation rules

The constraint grammar is composed of 299 rules which are divided into four categories: Safe, Safe heuristic, Heuristic, and Syntax labeling. The distribution of rules is shown in Table 2.

The philosophy is that Safe rules should represent real constraints in the language. Examples might be that a preposition cannot directly precede a finite verb or that prepositional case requires a preceding preposition.

[4] `http://dict.ruslang.ru/gram.php`

	SELECT	REMOVE	MAP
Safe	16	34	–
Safe heuristic	89	76	–
Heuristic	26	52	–
Syntax labelling	–	–	6

Table 2: The 299 rules in the grammar are separated into four sections depending on rule reliability.

Safe heuristic rules should deal with highly frequent tendencies in the language. For example remove a genitive at the beginning of a sentence if it is capitalised and there is no verb governing the genitive found to the right and there is also no negated verb to the right. This rule relies on the fact that if the genitive is in first position in the sentence it cannot modify anything before, and no preposition can be governing it. This kind of rule often relies on completeness of sets, in this case the set of verbs that can take a genitive complement.

Heuristic rules are those which we do not consider linguistic constraints, but express preferences, often dealing with overgeneration or overspecification in the morphological transducer. For example, remove the verbal adverb reading of такая, which could be the feminine singular nominative of такой 'such' or the verbal adverb of такать 'say *well...*'.

Given a large hand-annotated corpus we believe that most of the heuristic rules would be better replaced with information learnt from the corpus through stochastic methods.

5 Development process

A common approach taken when writing constraint grammar rules is to apply the existing rule set to a new text, write new rules to deal with the ambiguities, then apply the rules to a hand-annotated corpus to see how often the rule disambiguated correctly (Voutilainen, 2004).

Due to the lack of a hand-annotated corpus compatible with our morphological analyser, we adopted a slightly modified technique. We picked a random text from the Russian Wikipedia,[5] ran it through the morphological analyser, wrote rules, and then ran the rules on the whole Wikipedia corpus. For each rule, we collected around 100 example applications and checked them. If a rule selected the appropriate reading in all cases, we included it in the *safe* rule set, if it removed an appropriate reading in less then three cases, then we included it in the *safe heuristic* rule set. Otherwise we either discarded the rule or included it in the heuristic rule set.

[5]The Russian Wikipedia was chosen as a testing corpus as it is the largest, freely licensed corpus of Russian available on the internet. It is not representative of Russian texts as a whole.

6 Evaluation

6.1 Corpus

In order to evaluate the grammar we hand-annotated 10,150 words of Russian text from Wikipedia articles, public domain literature and freely-available news sources. The annotated texts are available online under the CC-BY-SA licence.[6]

Hand-annotation proceeded as follows: The text was first morphologically analysed, and then an annotator read through the output of the morphological analyser, commenting out the readings which were not appropriate in context. This annotated text was then checked by a second annotator.

We chose to annotate our own texts as opposed to using a well-known hand-annotated corpus such as the Russian National Corpus (RNC) for two main reasons: the first was that the RNC is not freely available; the second was that the standards for tokenisation, part-of-speech and morphological description are different from our morphological analyser.

Table 3 gives a quantitative evaluation of the performance of our CG on the test corpus.

6.2 Qualitative evaluation

In this section, we give a qualitative evaluation of errors made by the CG.

Bad linguistics: In some cases a rule did not take into account grammatical possibilities in the language. e.g. Two simple rules such as

- `REMOVE Det IF (0 Det OR Pron) (1C Ne) ;`
- `REMOVE Det IF (0 Det OR Pron) (1 Cm LINK 1 CC OR CS) ;`

did not take into account the possibility of having a postposed determiner as in

- ...а может быть и раньше, и факт этот не раз поражал меня ...

[6]https://svn.code.sf.net/p/apertium/svn/languages/apertium-rus/texts/

```
"<В>"
    "в" pr

"<ноябре>"
    "ноябрь" n m nn sg prp

"<1994>"
    "1994" num

"<года>"
    "год" n m nn sg gen SELECT:r462
;   "год" n m nn pl nom fac SELECT:r462

"<в>"
    "в" pr

"<Танзании>"
    "Танзания" np al f nn pl acc
    "Танзания" np al f nn sg prp
;   "Танзания" np al f nn pl nom REMOVE:r424
;   "Танзания" np al f nn sg dat REMOVE:r433
;   "Танзания" np al f nn sg gen REMOVE:r433

"<начал>"
    "начало" n nt nn pl gen
    "начать" vblex perf tv past m sg
;   "начать" vblex perf iv past m sg REMOVE:r769

"<работу>"
    "работа" n f nn sg acc

"<Международный>"
    "международный" adj m an sg nom
    "международный" adj m nn sg acc

"<трибунал>"
    "трибунал" n m nn sg acc
    "трибунал" n m nn sg nom

"<по>"
    "по" pr

"<Руанде>"
    "Руанда" np al f nn sg prp
    "Руанда" np al f nn sg dat
"<.>"
    "." sent
```

Figure 1: Example output from the morphological analyser and constraint grammar for the sentence В ноябре 1994 года в Танзании начал работу Международный трибунал по Руанде. "The work of the International Tribunal for Rwanda started in Tanzania in November 1994." The input ambiguity is 1.76 readings per word and the output ambiguity is 1.38 readings per word. Recall is 1.0 and precision is 0.72. Figure 2 shows the rules that fired for this example sentence.

```
### Safe

SELECT:r462 Gen IF (0 Year) (-1 Num LINK -1 Months LINK -1 Pr/V);
  # Select genitive reading of 'года' if there is a numeral immediately
  # to the left, before that there is a month and before that there is
  # the preposition 'в'.

REMOVE:r424 Nom IF (-1C Pr) ;
  # Remove nominative case if there is a word which can only be a
  # preposition immediately to the left.

REMOVE:r433 NGDAIP - Acc - Prp - Loc IF
               (-1C* Pr/V OR Pr/Na
                 BARRIER (*) - Adv - Comp - DetIndecl - ModAcc - ModPrp);
  # Remove all cases apart from accusative, preposition and locative
  # if 'в' or 'на' are found to the left and are unambiguous. The barrier
  # is anything that cannot be found inside a noun phrase.

### Safe heuristic

REMOVE:r769 IV IF (0 TV OR IV) (1C Acc) (NOT 1 AccAdv);
  # Remove an intransitive reading of a verb if the next word can only
  # be accusative and is not in the set of nouns which can be used
  # adverbially in accusative.
```

Figure 2: Some example rules from the grammar.

Domain	**Tokens**	**Precision**	**Recall**	**F-score**	**Ambig. solved**
Wikipedia	7,857	0.506	0.996	0.671	44.92%
Literature	1,652	0.473	0.984	0.638	42.95%
News	642	0.471	0.990	0.638	41.60%
Average	10,150	0.498	0.994	0.663	44.39%

Table 3: Results for the test corpora.

- . . . and maybe even earlier, and fact **this** not once surprised me . . .

or a interposed parenthetical as in

- Но какие, однако же, два разные создания, точно обе с двух разных планет!
- But what, **exactly** , two different creatures, just both from two different planets!

Bad rule: In some cases a rule was simply incorrectly specified. For example, the following rule was designed to solve the ambiguity between short-form neuter adjectives and adverbs

- `REMOVE A + Short IF (-1C Fin OR Adv OR A) (0C Short OR Adv) ;`

However there is no reason why we should prefer an adverb over an adjective after an adverb,

- . . . потому что совсем неприятно проснуться в гробу под землею.
- . . . because [it is] really **unpleasant** to wake up in a coffin under the ground.

Incomplete barrier: Some rules suffered from incomplete barriers, which is something that would benefit from a more systematic treatment.

- `REMOVE NGDAIP - Acc - Prp - Loc IF (-1C* Pr/V OR Pr/Na BARRIER (*) - Adv - Comp - DetIndecl - ModAcc - ModPrp) ;`

here the rule removes the nominative reading of the adjective to leave the accusative reading because the preposition в 'in' is found preceeding.

- В 1960-х электрифицированные высокоскоростные железные дороги появились в Японии и некоторых других странах.
- In the 1960's **electrified** high-speed railways appeared in Japan and some other countries.

Incomplete set: In some cases the rule was a good generalisation, but made use of a set which was incomplete. For example:

- `REMOVE Dat IF (NOT 0 Prn/Sebe) (NOT 0 Anim OR Cog OR Ant) (NOT 0 Pron) (NOT 1* V/Dat) (NOT -1* V/Dat) (NOT -1* Prep/Dat) (NOT -1C A + Dat) ;`

the set `V/Dat` does not contain the verb противопоставлять 'opposed to' which takes a dative argument.

- В связи с этим ортодоксальности стали противопоставлять ересь.
- In connection with this **orthodoxy** was opposed to heresy.

Rule interaction: The strong accusative rule below causes incorrect behaviour in the rule to remove transitivity readings

- `REMOVE TV - Pass IF (NOT 1* Acc) (NOT -1* Acc) ;`
- `REMOVE Acc IF (-1C Fin + IV) (NOT 0 AccAdv) ;`

Consider the following example where может 'can' is tagged as intransitive, the second rule fires removing the accusative reading of его 'him', and thus given the lack of accusative reading, найти 'find' is disambiguated as intransitive instead of transitive.

- Она смотрит везде, но не может его найти.
- She looks around, but she cannot **find him**.

Difficult linguistics: Dealing with participles with arguments is challenging in the case that the arguments of the participle share the same government as the main verb.

- `REMOVE IV IF (0 TV OR IV) (1C Acc) (NOT 1 AccAdv) ;`

Here Ваню и Машу 'Vanja and Maša' are the object of видит 'sees' and not играющих 'playing', although both verbs can take accusative object.

- Их мама внутри дома с кошкой, она смотрит в окно и видит играющих Ваню и Машу.

- Their mother is inside the house with the cat, she looks through the window and sees Vanja and Maša **playing**.

This kind of error would ideally be resolved with semantic knowledge.

6.3 Task-based evaluation

The constraint grammar described in this paper has been applied to the task of automatic word stress placement (Reynolds and Tyers, 2015). This task is especially relevant for Russian language learners, because vowels are pronounced differently depending on their position relative to stress position. For example, the word *molokó* 'milk' is pronounced /məlʌkɔ/, where each instance of the letter *o* corresponds to a different vowel sound. Russian has complicated patterns of shifting stress, which are difficult for learners to master. Almost 99% of wordforms with ambiguous stress position can be disambiguated morphosyntactically, so a constraint grammar can potentially resolve most stress ambiguity indirectly. The results of Reynolds and Tyers (2015) show that our constraint grammar overcomes about 42% of the ambiguity relevant to stress ambiguity in unrestricted text.

6.4 Combining with a statistical tagger

Given that just over half of all ambiguity remains after running our preliminary constraint grammar and that for many applications unambiguous output is necessary, we decided to experiment with combining the constraint grammar with a statistical tagger to resolve remaining ambiguity. Similar approaches have been taken by previous researchers with Basque (Ezeiza et al., 1998), Czech (Hajič et al., 2001; Hajič et al., 2007), Norwegian (Johannessen et al., 2011; Johannessen et al., 2012), Spanish (Hulden and Francom, 2012), and Turkish (Oflazer and Tür, 1996).

We follow the voting method described by Hulden and Francom (2012). We used the freely available `hunpos` part-of-speech tagger (Halácsy et al., 2007). We performed 10-fold cross validation using our evaluation corpus, taking 10% for testing and 90% for training, and experimented with three configurations:

- `HMM`: the `hunpos` part-of-speech tagger with its default options
- `HMM+Morph`: as with `HMM` but incorporating the output of our morphological analyser (see section 4.1) as a full form lexicon.
- `HMM+Morph+CG`: we submitted the output from `HMM+Morph` and the constraint grammar to a voting procedure, whereby if the constraint grammar left one valid reading, we chose that, otherwise if the constraint grammar left a word with more than one reading, we chose the result from the `HMM+Morph` tagger.

As can be seen from Figure 3, incorporating the constraint grammar improves the performance of the HMM tagger, an improvement of nearly 5% in accuracy, similar to that reported by Hulden and Francom (2012) for the same amount of training data. In Figure 3, it appears that the HMM alone is much more dependent on training corpus size than the voting setup, which improves very little between a training corpus size of 5,000 and 9,000.

Our constraint grammar also has a much lower precision as a result of the ambiguity remaining in the output. Similarly, the final accuracy is below the state of the art for Russian. For instance, Sharoff et al. (2008) report a maximum accuracy of 95.28% using the TnT tagger. Note, however, that this model was trained on a much larger corpus – over five million tokens – which is not freely available.

7 Future work

We have a number of plans for future work, the first of which is increasing the precision of the grammar without decreasing recall. Secondly we would like to add syntactic function labelling and dependency parsing. For the dependency parser we plan to reuse the Giellatekno dependency grammar as in (Antonsen et al., 2010).

The development workflow could also be improved, for example during the testing of each rule we could save the correct decisions of the grammar. This would give us a partially-disambiguated development corpus, which could be gradually used to build up a gold-standard corpus, and which could also be used for regression testing to ensure that new rules added do not invalidate the correct decisions of previously written rules.

Also it is worth noting that although Russian has a great deal of non-free resources, this paper also presents a method which is promising

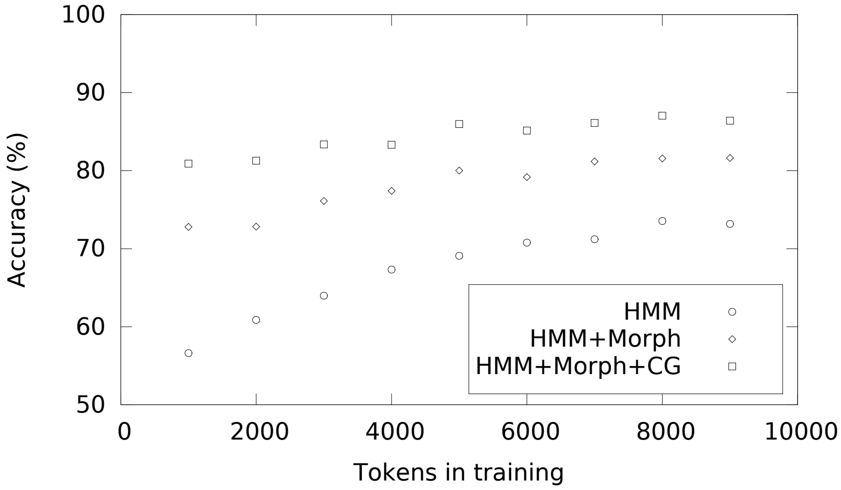

Figure 3: Learning curve for three taggers, `hunpos` with no lexicon, `hunpos` with a lexicon, and `hunpos` with a lexicon and the Russian constraint grammar in a voting set up.

for smaller or lesser-resourced Slavic languages such as Sorbian, Rusyn or Belarusian. Instead of hand-annotating a large quantity of text, it may be more efficient to work on grammatical resources — such as a morphological analyser and constraint grammar — and use them alongside a smaller quantity of high-quality annotated text.

8 Conclusions

This paper has presented a preliminary constraint grammar for Russian, where rules have been assigned to sections based on observations of performance on a non-gold corpus. The constraint grammar is high recall (over 0.99) and improves the performance over a trigram HMM-based tagger. It also shows state-of-the-art performance for the stress-placement task.

Acknowledgments

We are grateful to Koen Claessen for insightful discussion, as well as three anonymous reviewers who gave thoughtful feedback on an earlier version of this paper. All remaining errors are our own.

References

Lene Antonsen, Linda Wiechetek, and Trond Trosterud. 2010. Reusing grammatical resources for new languages. In *Proceedings of the International conference on Language Resources and Evaluation LREC2010*, pages 2782–2789.

Kenneth R Beesley and Lauri Karttunen. 2003. *Finite-state morphology: Xerox tools and techniques*. CLSI, Stanford.

Nerea Ezeiza, Iñaki Alegria, José María Arriola, Rubén Urizar, and Itziar Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 380–384. Association for Computational Linguistics.

Jan Hajič, Pavel Krbec, Pavel Květoň, Karel Oliva, and Vladimír Petkevič. 2001. Serial combination of rules and statistics: A case study in czech tagging. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 268–275. Association for Computational Linguistics.

Jan Hajič, Jan Votrubec, Pavel Krbec, Pavel Květoň, et al. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 67–74. Association for Computational Linguistics.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. Hunpos: An open-source trigram tagger. In *Proceedings of the 45th annual meeting of the ACL*, pages 209–212.

Mans Hulden and Jerid Francom. 2012. Boosting statistical tagger accuracy with simple rule-based grammars. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.

Janne Bondi Johannessen, Kristin Hagen, André Lynum, and Anders Nøklestad. 2011. OBT+Stat: Evaluation of a combined CG and statistical tagger. In Eckhard Bick, Kristin Hagen, Kaili Müürisep, and Trond Trosterud, editors, *Proceedings of the NODALIDA 2011 Workshop Constraint Grammar Applications*, volume 14, pages 26–34, Riga, Latvia. NEALT.

Janne Bondi Johannessen, Kristin Hagen, André Lynum, and Anders Nøklestad. 2012. Obt+stat: A combined rule-based and statistical tagger. In Gisle Andersen, editor, *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*, pages 51–66. John Benjamins Publishing.

Kimmo Koskenniemi. 1984. A general computational model for word-form recognition and production. In *Proceedings of the 10th International Conference on Computational Linguistics*, COLING '84, pages 178–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Krister Linden, Miikka Silfverberg, Erik Axelson, Sam Hardwick, and Tommi Pirinen. 2011. Hfst—framework for compiling and applying morphologies. In Cerstin Mahlow and Michael Pietrowski, editors, *Systems and Frameworks for Computational Morphology*, volume Vol. 100 of *Communications in Computer and Information Science*, pages 67–85. Springer.

Igor Nozhov. 2003. Морфологическая и синтаксическая обработка текста (модели и программы) *[Morphological and Syntactic Text Processing (models and programs)] also published as* Реализация автоматической синтаксической сегментации русского предложения *[Realization of automatic syntactic segmentation of the Russian sentence]*. Ph.D. thesis, Russian State University for the Humanities, Moscow.

Kemal Oflazer and Gökhan Tür. 1996. Combining hand-crafted rules and unsupervised learning in constraint-based morphological disambiguation. In *Proceedings of the ACLSIGDAT Conference on Empirical Methods in Natural Language Processing*, pages 69–81, Philadelphia, PA, USA.

Robert Reynolds and Francis Tyers. 2015. Automatic word stress annotation of Russian unrestricted text. In *Main conference proceedings from NODALIDA 2015*, Vilnius, Lithuania. NEALT.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *International Conference on Machine Learning; Models, Technologies and Applications*, pages 273–280.

Serge Sharoff, Mikhail Kopotev, Tomaž Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a Russian tagset. In *Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008*, Marrakech.

Atro Voutilainen. 2004. Hand crafted rules. In H. van Halteren, editor, *Syntactic Wordclass Tagging*, pages 217–246. Kluwer Academic.

Andrej Anatoljevič Zaliznjak. 1977. Грамматический словарь русского языка: словоизменение: около 100 000 слов *[Grammatical dictionary of the Russian language: Inflection: approx 100 000 words]*. Изд-во "Русский язык".