The Algorithmic Inflection and Morphological Variability of Russian

Anonymous ACL submission

Abstract

We present a set of deterministic algorithms for Russian inflection and automated text synthesis. These algorithms are implemented in a publicly available web-service www.passare.ru. This service provides functions for inflection of single words, word matching and synthesis of grammatically correct Russian text. Selected code and datasets are available at https://github.com/passare-ru/PassareFunctions/ Performance of the inflectional functions has been tested against the annotated corpus of Russian language `OpenCorpora` (OpenCorpora), compared with that of other solutions, and used for estimating the morphological variability and complexity of different parts of speech in Russian.

1 Introduction

Automatic inflection of words in a natural language is necessary for a variety of theoretical and applied purposes like parsing, topic-to-question generation (Chali and Hasan, 2015), speech recognition and synthesis, machine translation (Streiter et al., 2000), tagset design (Kuzmenko, 2017), information retrieval (Iomdin, 2003), content analysis (Belonogov et al., 2010; Belonogov and Kotov, 1971), and natural language generation (Cerutti et al., 2019; Costa et al., 2018; Subramanian et al., 2017; Tran et al., 2017). Various approaches towards automated inflection have been used to deal with particular aspects of inflection (Conway, 2001; Zaliznyak, 1967) in predefined languages (Foust, 1960; Fuks, 2010; Raja et al., 2014; Korobov, 2015; Porter, 1980) or in an unspecified inflected language (Faruqui et al., 2015; Silberztein, 2016).

Despite substantial recent progress in the field (d'Ascoli et al., 2020; Buddana et al., 2021; Korobov, 2015; Silberztein, 2016; Sorokin, 2016; Xiao et al., 2013), automatic inflection and automatic text generation still represent a problem of formidable computational complexity for many natural languages in the world. Most state-of-the-art approaches make use of extensive manually annotated corpora that currently exist for all major languages (Segalovich, 2003). Real-time handling of a dictionary that contains millions of inflected word forms and tens of millions of relations between them is not an easy task (Goldsmith, 2001). Besides, no dictionary can ever be complete. For these reasons, algorithmic coverage of the grammar of a natural language is important provided that inflection in this language is complex enough.

Russian is a highly inflected language whose grammar is known for its complexity (Sorokin, 2016; Zaliznyak, 1967). In Russian, inflection of a word may require changing its prefix, root, and ending simultaneously while the rules of inflection are highly complex (Halle and Matushansky, 2006; Zaliznyak, 1967). The form of a word can depend on as many as seven grammatical categories such as number, gender, person, tense, case, voice, animacy etc (cf Fig. 1). By an estimate based on (OpenCorpora), the average number of different grammatical forms of a Russian adjective is 11.716. A Russian verb has, on average, 44.069 different inflected forms, counting participles of all kinds and the gerunds (cf. Fig. 1).

In the present paper we describe a fully algorithmic dictionary-free approach towards automatic inflection of Russian. The algorithms exposed in the paper are based on morphological rules that have been manually created and tuned by the authors. Implementation of the algorithms has been performed in C# programming language, see https://github.com/passare-ru/PassareFunctions/ The described functionality is freely available online at www.passare.ru through both manual entry of a word to be inflected and by API access of main functions for dealing with big amounts of data.

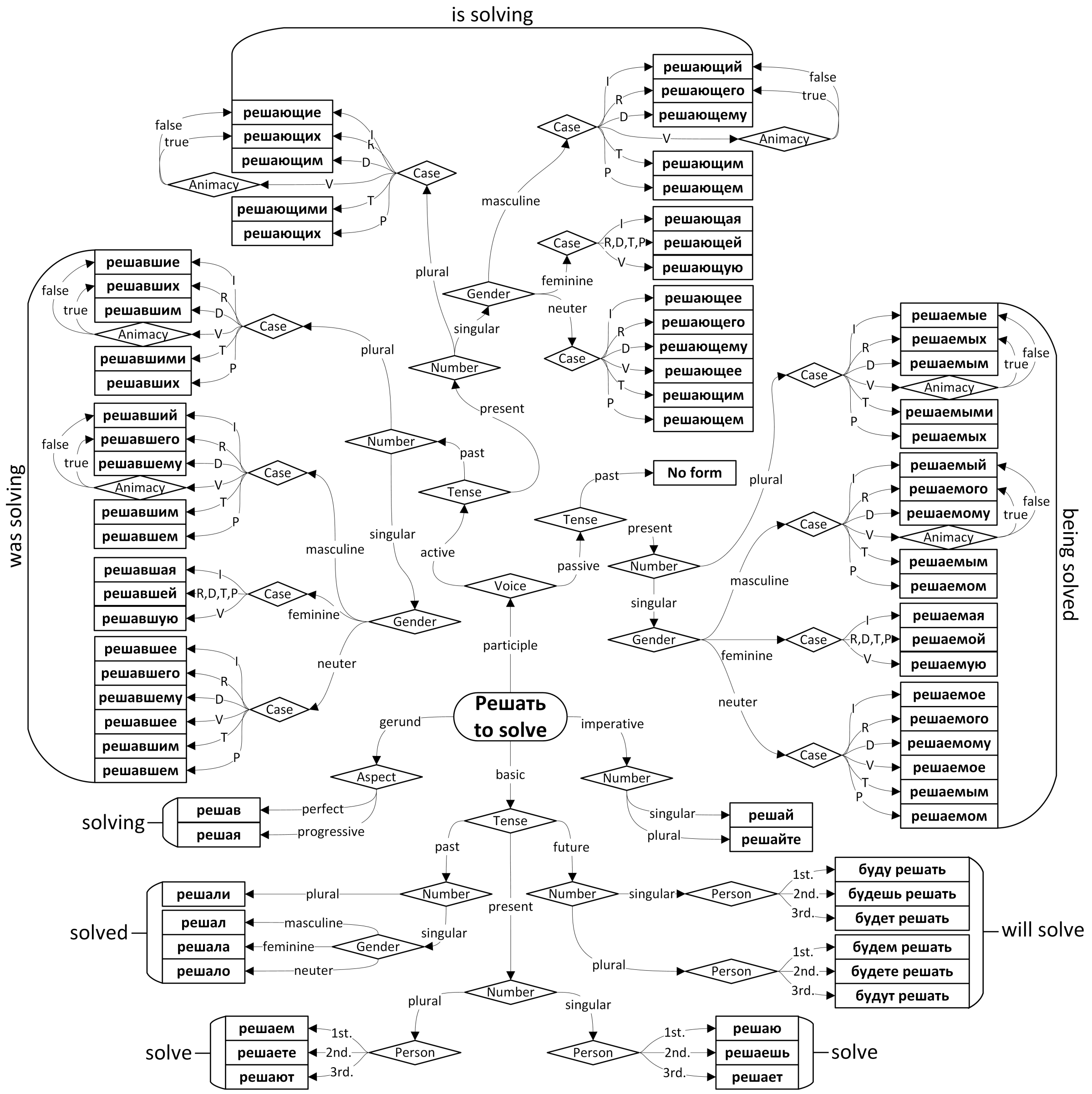

Figure 1: All of the forms of the Russian verb "решать" – "reshat′ " – "to solve" and their dependence on tense (past, present, or future), number (singular or plural), gender (feminine, masculine, or neuter), person (1st, 2nd, or 3rd), voice (active or passive), aspect (perfect or progressive), case (nominative, genitive, dative, accusative, instrumental, or prepositional, abbreviated in accordance with Russian translation by I, R, D, V, T, and P, respectively), and animacy (boolean-valued)

2 Inflection in Russian Language: Algorithms and Implementation

The web-service passare.ru offers a variety of functions for inflection of single Russian words, word matching, and synthesis of grammatically correct text. In particular, the inflection of a Russian noun by number and case, the inflection of a Russian adjective by number, gender, and case, the inflection of a Russian adverb by the degrees of comparison are implemented. Russian verb is the part of speech whose inflection is by far the most complicated in the language. The implemented algorithms provide inflection of a Russian verb by tense, person, number, and gender. These algorithms also allow one to form the gerunds and the imperative forms of a verb. Besides, functions for forming and inflecting active present and past participles as well as passive past participles have been implemented. Passive present participle is the only verb form not currently supported by the website due to the extreme level of its irregularity. Besides, passive present participle cannot be

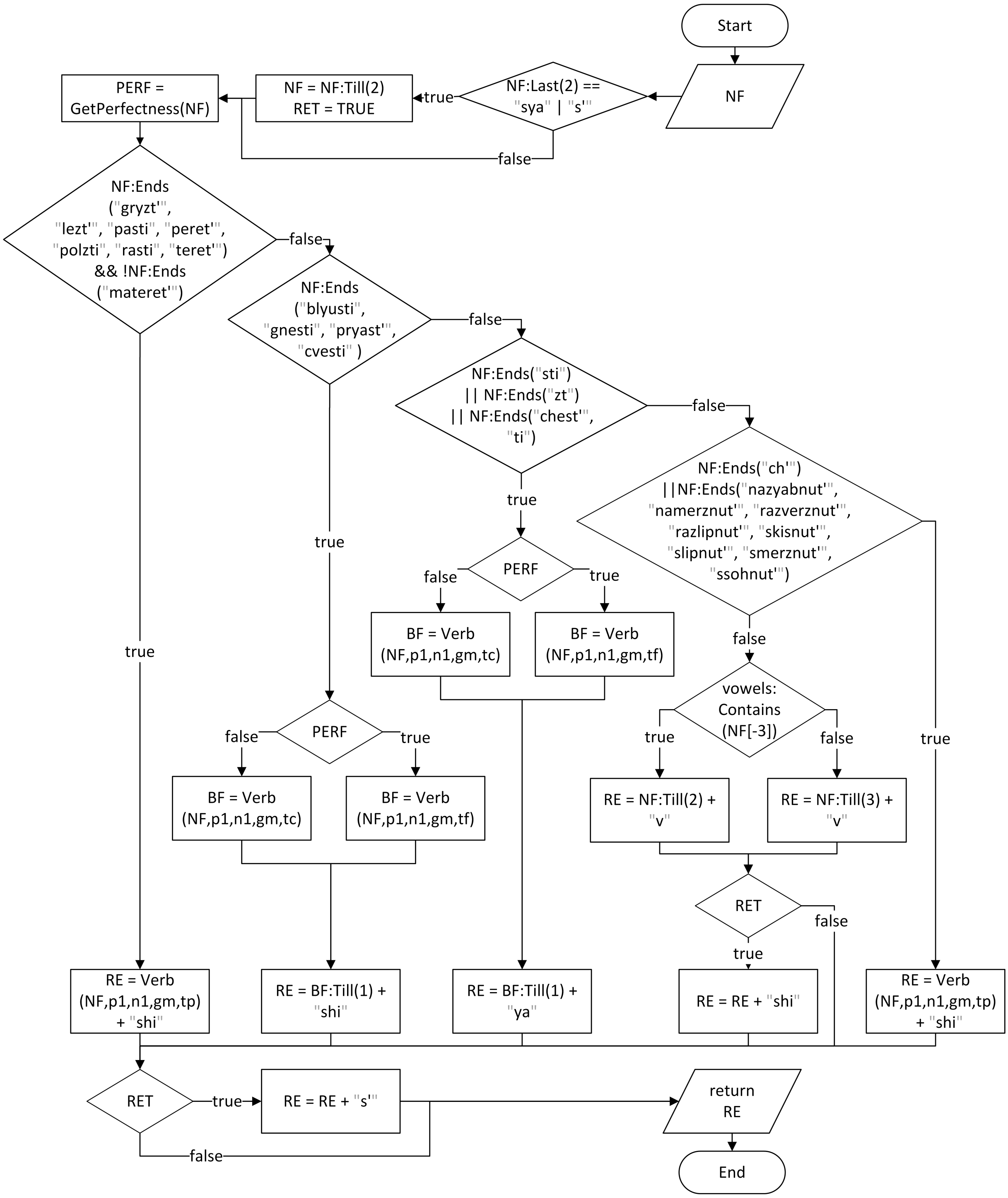

Figure 2: Generation of the perfective gerund form of a verb. Russian words and letters are given in conventional English transliteration. `NF`, `BF`, and `RE` stand for the normal form of a verb, the base of a verb, and the result of the computation, respectively. `PERF` and `RET` are boolean variables encoding perfective and reflexive properties of a verb, respectively. The list `vowels` comprises Russian vowels. The other notation coincides with the C# syntax, NF:Till(2) standing for the string NF will two last characters removed. The corresponding C# code is available at https://github.com/passare-ru/PassareFunctions/

formed at all for numerous verbs in the Russian language.

The algorithmic coverage of the Russian language provided by the web-service passare.ru aims to balance grammatical accuracy and ease of use. For that reason, a few simplifying assumptions have been made: the Russian letters "ë" and "e" are treated as identical; no information on the stress in a word is required to produce its inflected forms; for inflectional functions, the existence of an input word in the language is determined by the user. Furthermore, the animacy of a noun is not treated as a variable category in the noun-inflecting function despite the existence of 1037 nouns (about 1.4% of the nouns in the `OpenCorpora` database (OpenCorpora)) with unspecified animacy. This list of nouns has been manually reviewed by the authors on a case-by-case basis and the decision has been made in favor of the form that is more frequent in the language than the others. The other form can be obtained by calling the same function with a different `case` parameter (`Nominative` or `Genitive` instead of `Accusative`).

Similarly, the perfectiveness of a verb has not been implemented as a parameter in a verb-inflecting function although by (OpenCorpora) there exist 1038 verbs (about 3.2% of the verbs in the database) in the language whose perfectiveness is not specified. For such verbs, the function produces forms that correspond to both perfective and imperfective inflections.

The inflectional form of a Russian word defined by a choice of grammatical categories (such as number, gender, person, tense, case, voice, animacy etc.) is in general not uniquely defined. This applies in particular to many feminine nouns, feminine forms of adjectives, and to numerous verbs. For such words, the algorithms implemented in the web-service passare.ru only aim at finding one of the inflectional forms, typically, the one which is the most common in the language.

Due to the rich morphology of the Russian language and to the high complexity of its grammar, a detailed description of the algorithms of Russian inflection cannot be provided in a short research paper. The algorithm for the generation of the perfective gerund form of a verb is presented in Fig. 2. Using the verb "решать" – "reshat′" – "to solve" as input, the algorithm outputs the gerund "решав" – "reshav" – "having solved". Most of the notation in Fig. 2 is the same as that of the C# programming language. Furthermore, `NF` denotes the input normal form (the infinitive) of a verb to be processed. `GetPerfectness()` is a boolean function which detects whether a verb is perfective or not. `Verb()` is the function which inflects a given verb with respect to person, number, gender and tense. `BF` denotes the basic form of a Russian verb which is most suitable for constructing the perfective gerund of that verb. We found it convenient to use one of the three different basic forms depending on the type of the input verb to be inflected. The list `vowels` comprises all vowels in the Russian alphabet.

Although Russian morphology is extensively covered in the literature, the algorithms of the web-service www.passare.ru are in general fully novel and very different from other existing algorithmic approaches or textbook rules. The implementation comprises about 35,000 lines of code and has been compiled into a 571 kB executable file.

3 Software Speed Tests and Verification of Results

The software being presented has been tested against one of the largest publicly available corpora of Russian, `OpenCorpora` (OpenCorpora). We have been using Intel Core i5-2320 processor clocked at 3.00GHz with 16GB RAM under Windows 10. With all indefinite forms of the words in the `OpenCorpora` database as input, the whole output produced by www.passare.ru has been checked against the corresponding forms in the database to see how many discrepancies are present. The results are summarized in Table 1.

All of the words whose inflected forms did not show full agreement with the `OpenCorpora` database have been manually reviewed by the authors on a case-by-case basis. In the case of nouns, 26.76% of all error-producing input words belong to the class of Russian nouns whose animacy cannot be determined outside the context (e.g. "ёж" – "yozh" – "a hedgehog" *or, depending on context,* "a Czech hedgehog", "жучок" – "zhuchok" – "a bug" *or, depending on context,* "a hidden microphone" and the like). For verbs, 11.26% of the discrepancies result from the verbs whose perfectiveness cannot be determined outside the context without additional information on the stress in the word (e.g. "насыпать" – "nasypat′" – "to pour on", "пахнуть" – "pakhnut′" – "to smell" *or, depending on the stress,* "to smack" etc.).

Besides, a number of errors in `OpenCorpora` have been discovered. The classification of flaws in `OpenCorpora` is beyond the scope of the present work and we only mention that the inflection of the verb "застелить" – "zastelit′" – "to cover" as well as the gerund forms of the

Table 1: Inflection speed and agreement rates of passare.ru and `OpenCorpora`

Part of speech	Total number of words	Total processing time, min:sec	Number of forms computed (per word)	Processing time per word, msec	Agreement rate with `OpenCorpora`
Noun	74633	02:36	12	2	98.557 %
Verb	32358	05:49	24	10	98.678 %
Adjective	42920	00:06	28	0.14	98.489 %
Adverb	1507	<00:01	2	0.021	n/a
Ordinal	10000 (range 0-9999)	00:30	18	3	n/a
Cardinal	10000 (range 0-9999)	00:23	24	2	n/a
Present participle active	16946	04:55	28	17	98.961 %
Past participle active	32358	10:19	28	19	99.152 %
Past participle passive	32358	10:32	28	19	94.803 %
Gerunds	32358	00:23	2	0.72	99.157 %
Verb imperative	32358	00:42	2	1	95.327 %

verbs "выместить" – "vymestit′" – "to take revenge on", "напечь" – "napech′" – "to bake", and "перекиснуть" – "perekisnut′" – "to go fully sour" appear to be incorrect in this database at the time of writing. In addition, certain gerunds of a class of reflexive verbs appear to be incorreclty listed in the database.

We remark that the average time needed for the generation of all inflected forms of an adjective is more than ten times shorter than that of a noun despite the fact that the number of forms of an adjective is greater. This fact reflects the high morphological regularity of adjectives in the Russian language whose exceptional inflection is primarily found within a class of possessive adjectives stemming from animated nouns.

Using the basic functions described above, one can implement automated synthesis of grammatically correct Russian text on the basis of any logical, numerical, financial, factual or any other precise data. The website passare.ru provides examples of such metafunctions that generate grammatically correct weather forecast and currency exchange rates report on the basis of real-time data available online. Besides, it offers a function that converts a correct arithmetic formula into Russian text.

Matching adjectives to nouns by gender and number, matching verbs to personal pronouns by person, gender, and number as well as numerous similar functions are implemented in the synthesis section of the website. These functions can also be used to put the components of a sentence into the grammatically correct forms.

4 Quantitative Corpus Analysis of Russian Morphological Complexity

We now use the algorithms implemented in the web-service www.passare.ru to analyze the complexity of inflection of different parts of speech in the Russian language. There are only three parts of speech that are of interest in this respect, namely, adjectives, nouns, and verbs (together with participles of all kinds). All other parts of speech in the Russian language either comprise a very limited number of words and their forms (like personal and possessive pronouns, conjunctions, interjections etc) or exhibit highly regular inflection (like adverbs). None of these parts of speech are interesting from the algorithmic inflection viewpoint since their irregular inflectional forms are very few and can be easily listed. On the contrary, inflection of adjectives, nouns and verbs in the Russian language is highly complex and often irregular (see Fig. 1 for verbs).

To measure the morphological variability of a word w we introduce the function

$$\mathcal{L}(w) := \sum_{i,j} \mathrm{dist}_L(w_i, w_j), \qquad (1)$$

where $\{w_i\}$ is the list of all forms of the word w (with a fixed order of values of grammatical parameters encoding these forms) and dist_L is the Levenshtein distance (Levenshtein, 1966) between the forms w_i and w_j. For instance, for the verb w := "решать" – "reshat′" – "to solve" the list $\{w_i\}$ of its forms comprises the 78 forms given in Fig. 1.

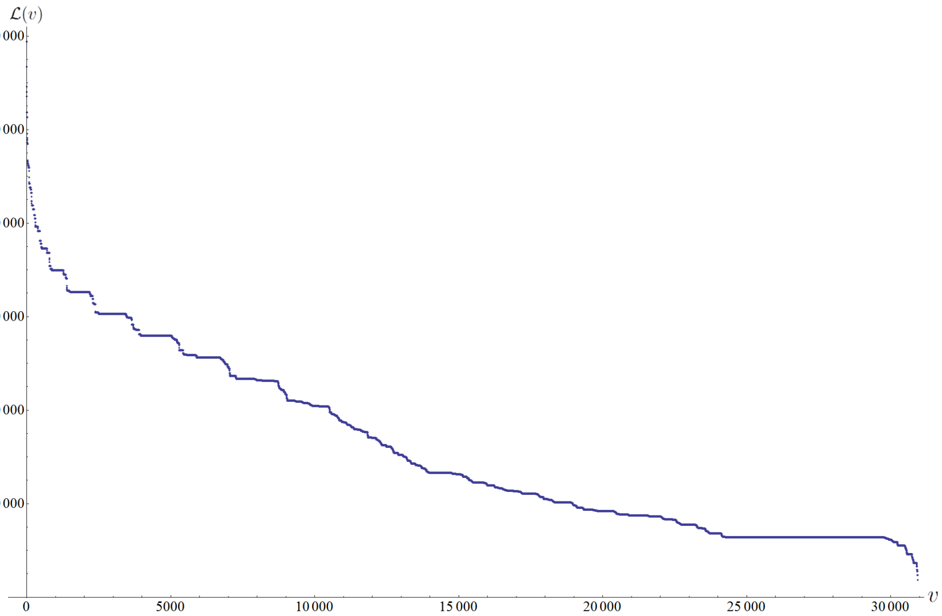

Figure 3: Morphological variability of verbs in the Russian language, verbs sorted by the values of $\mathcal{L}(v)$, the total Levenshtein distance (1) between the inflected forms of a verb v.

Verbs. Verbs exhibit the highest morphological variability among all parts of speech in the Russian language (cf Fig. 1). The algorithms for the inflection of verbs and producing various verb forms (participles and gerunds) are among the most complex in Russian grammar. Fig. 3 reflects the morphological variability of verbs in the Russian langauge. The horizontal axes corresponds to the 32358 Russian verbs listed in the `OpenCorpora` database. The height $\mathcal{L}(v)$ of a vertical segment corresponding to a verb v has been computed by means of the formula (1). The forms of a verb have been computed by means of the inflectional algorithms implemented at www.passare.ru. The performed analysis allows one to detect the Russian verbs (in the `OpenCorpora` database) with the extreme values of their inflectional variability. The overall shape of the monotone curve in Fig. 3 with few flat parts reflects the morphological complexity and very different inflectional patterns of verbs in Russian language. The vast majority of all verbs in the database (more precisely, 69.1% by our estimates) require detailed case analysis which has been performed in the algorithms implemented in the webservice passare.ru.

Adjectives. Adjectives are the part of speech with the most regular inflection in the Russian language. (Here we do not take into account parts of speech with very few words like personal pronouns, interjections, and the like.) Nevertheless, algorithmic inflection of Russian adjectives represents a task of substantial computational complexity.

Nouns. In Russian, nouns exhibit intermediate inflectional complexity compared to adjectives and verbs. Despite the vast majority of regular cases, there exist numerous exceptions which include e.g. indeclinable nouns of foreign origin.

A similar study has been carried out for other parts of speech in the Russian language which has led to a number of improvements in the inflectional algorithms.

5 Discussion

There exist several other approaches towards automated Russian inflection and synthesis of grammatically correct Russian text, e.g. (Kanovich and Shalyapina, 1994; Korobov, 2015). Besides, numerous programs attempt automated inflection of a particular part of speech or synthesis of a document with a rigid predefined structure (Chernikov and Karminsky, 2014). Judging by publicly available information, most of such programs make extensive use of manually annotated corpora which might cause failure when the word to be inflected is different enough from the elements in the database. The results of comparison of the approach exposed in the present paper with the other

Table 2: Comparing NLP software that offer Russian inflection or lemmatization

Software environment	Functionality	Supported languages	Dependency on dictionary	Distributed as	Implementation
passare.ru	inflection, word matching, data to text	Russian	low	free web service	algorithm extraction from language
morpher.ru	inflection (Nouns, Numerals), simple sentence matching	Russian, Ukrainian	high	commercial web service / standalone libraries	dictionary look-up
phpmorphy	morphological analysis, lemmatization, inflection	English, Russian, German, Ukrainian, Estonian and other	high	library (php)	dictionary look-up
pymorphy2	morphological analysis, lemmatization, inflection	Russian, Ukrainian	high	library (python)	dictionary look-up
NooJ	grammar development environment, linguistic analysis	arbitrary	high	framework	grammar based production
MARu	morphological analysis, lemmatization (using pymorphy2)	Russian	high, through pymorphy2 lemmatization	library (python)	various machine learning methods: linear model, CRF, deep neural network
natasha	segmentation, embeddings, morphology, lemmatization, syntax, NER, fact extraction	Russian	training data dependency, trained neural models dependency	several libraries (python)	`razdel` and `yargy` are rule-based systems; `navec` and `slovnet` are neural networks

software environments that offer functionality for Russian inflection or lemmatization are summarized in Table 2.

The speed of the computationally most expensive inflectional functions of www.passare.ru has been tested against that of the freeware products `phpmorphy` and `pymorphy2` on nouns, verbs and adjectives. The corresponding computation times on our system are 3:21, 4:13, and 7:06 for `phpmorphy` and 1:00, 2:44, and 1:58 `pymorphy2` (in min:sec format).

The solution presented in this paper has been designed to be as independent of any dictionary data as possible. However, due to numerous irregularities in the Russian language, several lists of exceptional linguistic objects (like the list of indeclinable nouns or the list of verbs with strongly irregular gerund forms, see Fig. 2) have been composed by the authors and used throughout the code, see https://github.com/passare-ru/PassareFunctions/ Whenever possible, rational descriptions of exceptional cases have been adopted to keep the numbers of elements in such lists to the minimum.

Acknowledgements

This research was performed in the framework of the state task in the field of scientific activity of the Ministry of Science and Higher Education of the Russian Federation, project no. FSSW-2020-0008.

References

G.G. Belonogov, A.A. Horoshilov, and A.A. Horoshilov. 2010. Automation of the English-Russian bilingual phraseological dictionaries based on arrays of bilingual texts. *Automatic*

Documentation and Mathematical Linguistics, 44(3):103–110.

G.G. Belonogov and R.G. Kotov. 1971. *Automated Information-Retrieval Systems*. Mir Publishers.

H.V.K.S. Buddana, S.S. Kaushik, P.V.S. Manogna, and S.K. P.s. 2021. Word level lstm and recurrent neural network for automatic text generation. *2021 International Conference on Computer Communication and Informatics, ICCCI 2021*.

F. Cerutti, A. Toniolo, and T.J. Norman. 2019. On natural language generation of formal argumentation. *CEUR Workshop Proceedings*, 2528:15–29.

Y. Chali and S.A. Hasan. 2015. Towards topic-to-question generation. *Computational Linguistics*, 41(1):1–20.

B.V. Chernikov and A.M. Karminsky. 2014. Specificities of lexicological synthesis of text documents. *Procedia Computer Science*, 31:431–439.

D. Conway. 2001. An algorithmic approach to English pluralization. In *Second Annual Perl Conference*. COPE.

F. Costa, P. Dolog, S. Ouyang, and A. Lawlor. 2018. Automatic generation of natural language explanations. *International Conference on Intelligent User Interfaces, Proceedings IUI*.

S. d'Ascoli, A. Coucke, F. Caltagirone, A. Caulier, and M. Lelarge. 2020. Conditioned text generation with transfer for closed-domain dialogue systems. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12379 LNAI:23–34.

M. Faruqui, Yu. Tsvetkov, G. Neubig, and C. Dyer. 2015. Morphological inflection generation using character sequence to sequence learning. *CoRR*, abs/1512.06110.

W.D. Foust. 1960. Automatic English inflection. In *National Symposium on Machine Translation*, pages 229–233. UCLA.

H. Fuks. 2010. Inflection system of a language as a complex network. *CoRR*, abs/1007.1025.

J. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

M. Halle and O. Matushansky. 2006. The morphophonology of Russian adjectival inflection. *Linguistic Inquiry*, 37(3):351–404.

L.L. Iomdin. 2003. Natural language processing as a source of linguistic knowledge. pages 68–74. Cited By 0.

M.I. Kanovich and Z.M. Shalyapina. 1994. The RUMORS system of Russian synthesis. *COLING*, pages 177–179.

M. Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. *Communications in Computer and Information Science*, 542:330–342.

E. Kuzmenko. 2017. Morphological analysis for Russian: Integration and comparison of taggers. *Communications in Computer and Information Science*, 661:162–171.

V.I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

OpenCorpora. An open corpus of Russian language.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

S.V.K. Raja, V. Rajitha, and M. Lakshmanan. 2014. Computational model to generate case-inflected forms of masculine nouns for word search in Sanskrit e-text. *J. Comput. Sci.*, 10(11):2260–2268.

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

M. Silberztein. 2016. *Formalizing Natural Languages: The NooJ Approach*. John Wiley and Sons Limited.

A. Sorokin. 2016. Using longest common subsequence and character models to predict word forms. *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, SIGMORPHON 2016 at the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 54–61.

O. Streiter, L. Iomdin, and I. Sagalova. 2000. Learning lessons from bilingual corpora: Benefits for machine translation. *International Journal of Corpus Linguistics*, 5(2):199–230.

S. Subramanian, S. Rajeswar, F. Dutil, C. Pal, and A. Courville. 2017. Adversarial generation of natural language. *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP 2017 at the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 241–251.

V.-K. Tran, L.-M. Nguyen, and S. Tojo. 2017. Neural-based natural language generation in dialogue using rnn encoder-decoder with semantic aggregation. *SIGDIAL 2017 - 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, pages 231–240.

T. Xiao, J. Zhu, and T. Liu. 2013. Bagging and boosting statistical machine translation systems. *Artificial Intelligence*, 195:496–527.

A.A. Zaliznyak. 1967. *Russian Nominal Inflection (Russian)*. Nauka Publishers.