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. The inflectional functions have been tested against the annotated corpus of Russian language `OpenCorpora` (OpenCorpora) 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 and Iomdin, 2000), tagset design (Kuzmenko, 2016), information retrieval (Iomdin, 1960), content analysis (Belonogov et al., 2010; Belonogov and Kotov, 1971), and natural language generation (Cerutti et al., 2017; Costa et al., 2017; Rajeswar et al., 2017; Tran and Nguyen, 2017). Various approaches towards automated inflection have been used to deal with particular aspects of inflection (William D, 2001; Ando and Zhang, 1967) in predefined languages (William D, 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 (Korobov, 2015; Silberztein, 2016; Sorokin, 2016; Xiao et al., 2013), automatic inflection still represents 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; Ando and Zhang, 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; Ando and Zhang, 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 described in the present paper are implemented in C# programming language. 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.

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

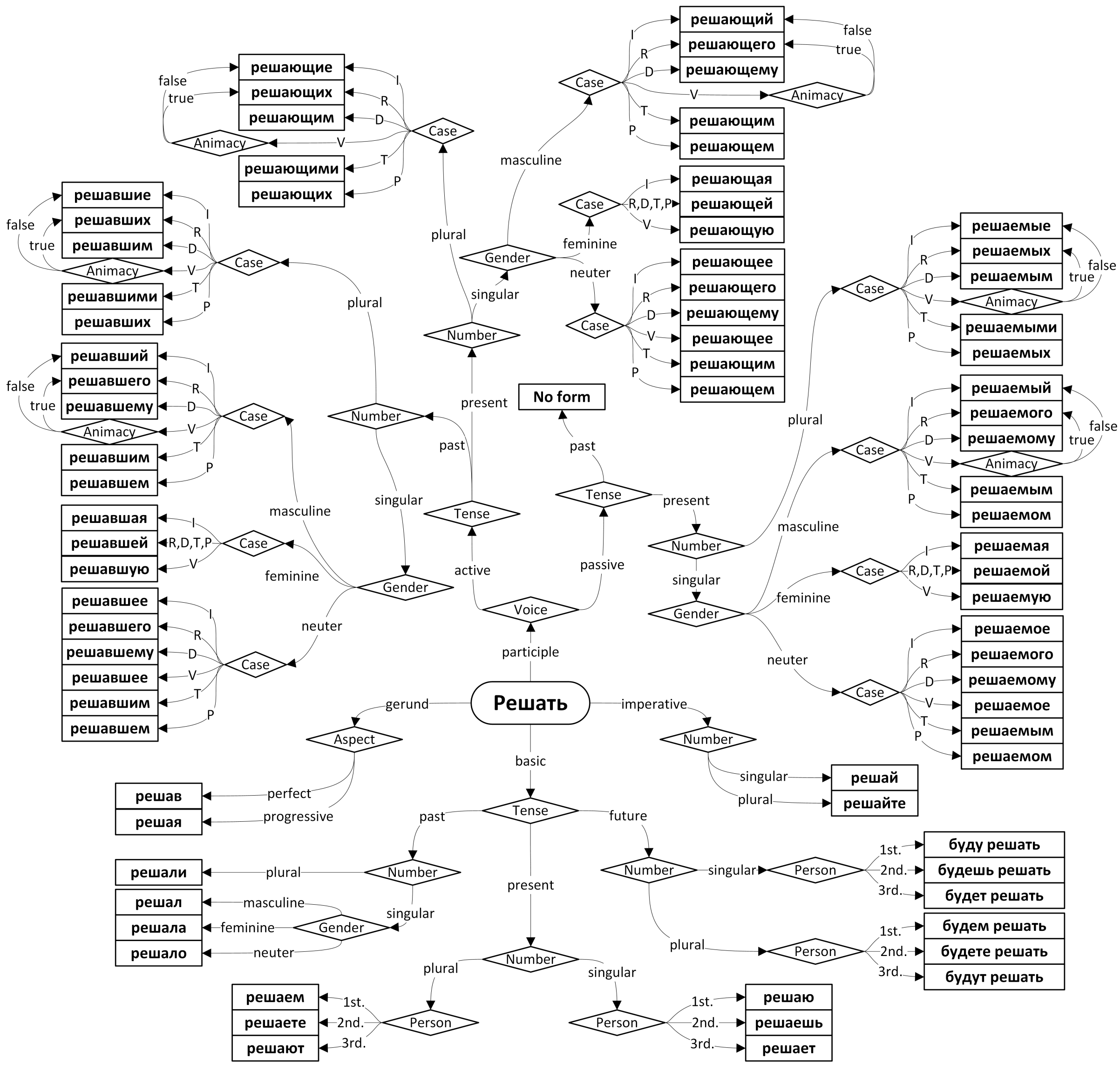

Figure 1: All of the forms of the Russian verb "решать" – "reshat′" – "to solve" and their dependence on tense, number, gender, person, voice, aspect, and case

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 presented algorithm provides inflection of a Russian verb by tense, person, number, and gender. It also allows 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 are implemented. Passive present participle is the only verb form not currently supported by the website due to the extreme level of its irregularity and absence for numerous verbs in the language.

The algorithmic coverage of the Russian language provided by the web-service passare.ru aims to balance grammatical accuracy and easiness of use. For that reason, a few simplifying assumptions have been made: the Russian letters "ё" and "е" are identified; 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

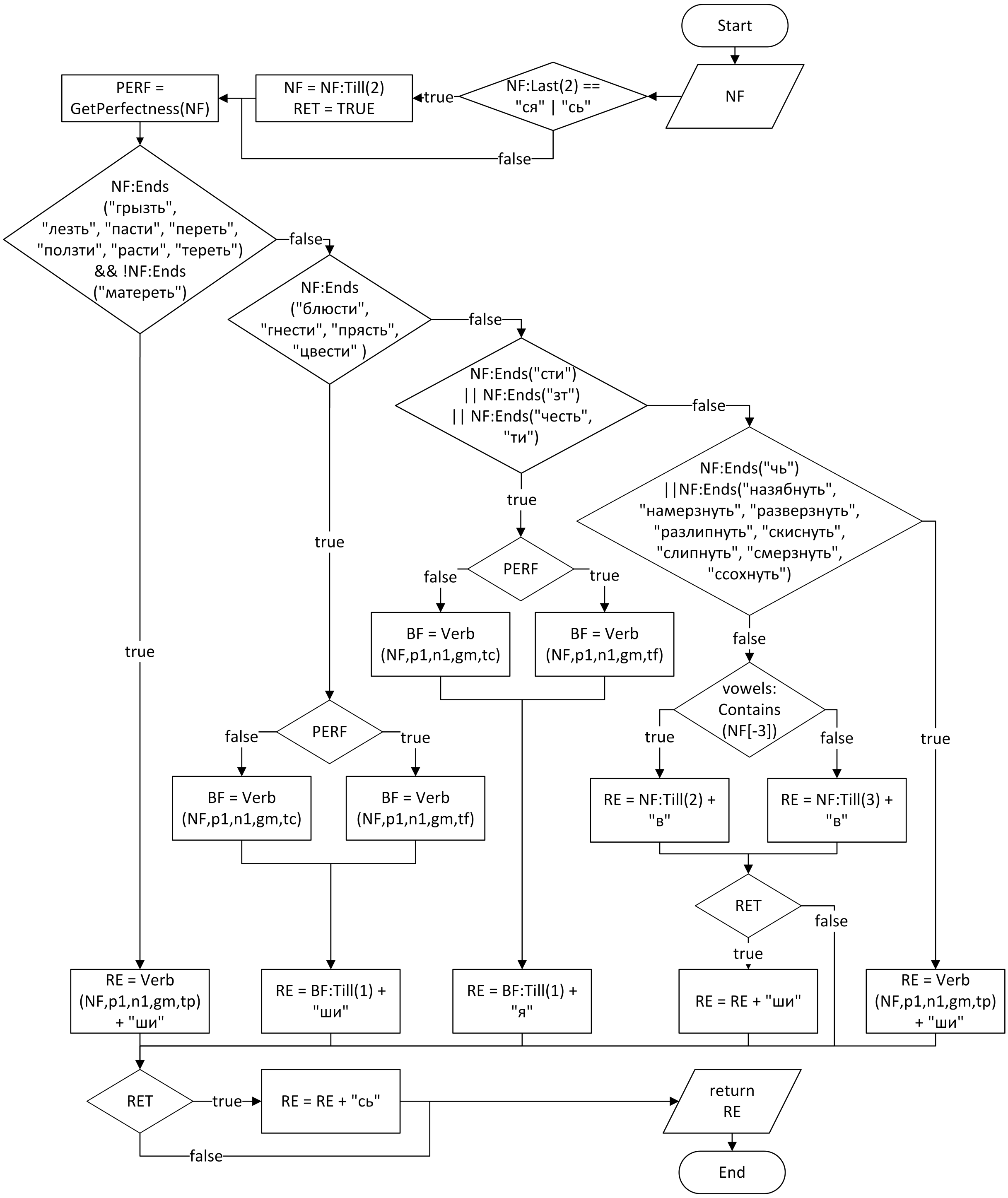

Figure 2: Generation of the perfective gerund form of a verb

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 is not 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, ani-

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 %

macy 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. 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. `GetPerfectiveness()` 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.

The algorithms have been implemented in C# programming language. 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. 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", "жучок" – "zhuchok" 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′", "пахнуть" – "pakhnut′" etc.).

Besides, a substantial 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′" as well as the gerund forms of the verbs "выместить" and "напечь" – "napech′" appear to be incorrect in this database at the time of writing.

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 pre-

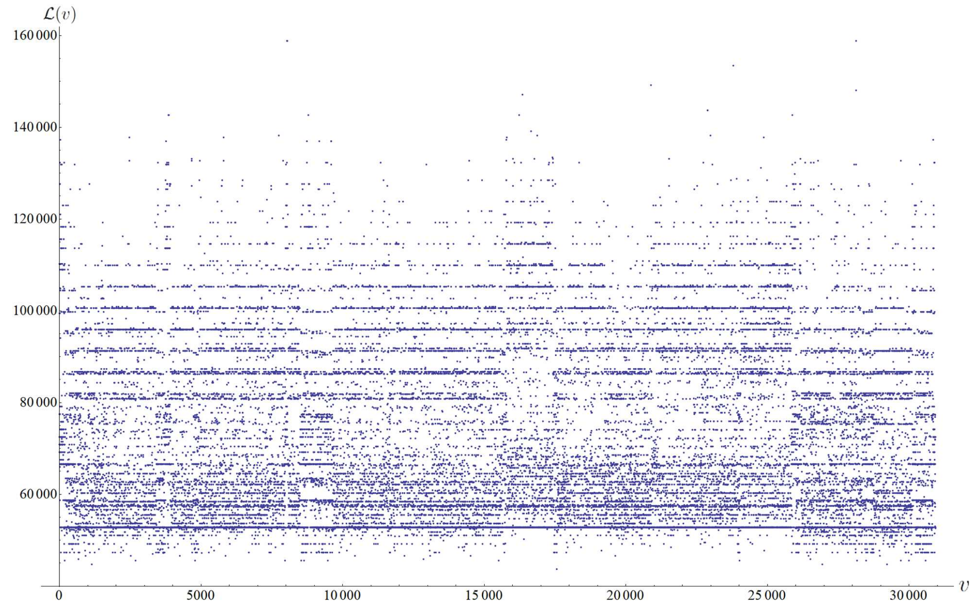

Figure 3: Morphological variability of verbs in the Russian language, verbs listed alphabetically

`Input:` (до ближайший среда (`_SINGULAR _PAST _3P _NGEN` остаться) (`_ICASE _CARDINAL` 2 день) (преобразовать в предложение))
`Output:` До ближайшей среды осталось два дня.

cise 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 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 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 if 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.

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

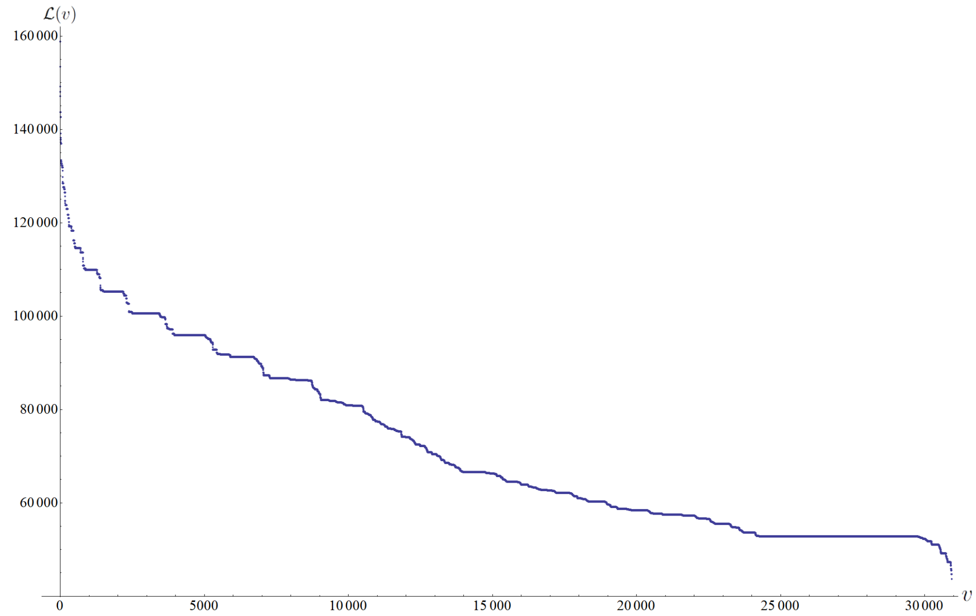

Figure 4: Morphological variability of verbs in the Russian language, verbs sorted by the values of $\mathcal{L}(v)$

has been computed by means of the formula (1). In this formula, $\{w_i\}$ is the list of all forms of a verb (with a fixed order of values of grammatical parameters encoding these forms) and dist_L is the Levenshtein distance (Levenshtein, 1966) between the verb forms w_i and w_j. 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. To emphasize the complexity of Russian verbal inflection we graph the function $\mathcal{L}(v)$ over the set of verbs sorted by the values of $\mathcal{L}$, see Fig. 4.

Each of the horizontal parts of the curve in Fig. 4 corresponds to a class of verbs whose inflection is described by a single rule with no exceptions. Together, they only represent 30.9% of the verbs in the `OpenCorpora` database. The remaining 69.1% of the verbs require detailed case analysis which has been performed in the algorithms described above.

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. The performed analysis of morphological variability of Russian adjectives is summarized in Fig. 5. Apart from the regular inflection patterns represented by the lines (1),(2), and (3) in Fig. 5, there exist numerous irregular adjectives that are almost uniformly distributed in the dictionary.

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 solution presented in this paper has been designed to be as independent of any dictionary

Figure 5: Morphological variability of adjectives in the Russian language, adjectives listed alphabetically

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 and used throughout the code. 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

Rie Kubota Ando and Tong Zhang. 1967. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

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

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

Federico Cerutti, Alice Toniolo, and Timothy J. Norman. 2017. On natural language generation of formal argumentation.

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

Boris V. Chernikov and Aleksander M. Karminsky. 2014. Specificities of lexicological synthesis of text documents. In *ITQM*, volume 31 of *Procedia Computer Science*, pages 431–439. Elsevier.

Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2017. Automatic generation of natural language explanations.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2015. Morphological inflection generation using character sequence to sequence learning. *CoRR*, abs/1512.06110.

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

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

Morris Halle and Ora Matushansky. 2006. The morphophonology of russian adjectival inflection. *Linguistic Inquiry*, 37(3):351–404.

Leonid L. Iomdin. 1960. Automatic english inflection. In *MLMTA*, pages 68–74. CSREA Press.

Max I. Kanovich and Zoya M. Shalyapina. 1994. The rumors system of russian synthesis. In *COLING*, pages 177–179.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *AIST*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer.

Elizaveta Kuzmenko. 2016. Morphological analysis for russian: Integration and comparison of taggers. In *AIST*, volume 661 of *Communications in Computer and Information Science*, pages 162–171.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

OpenCorpora. An open corpus of Russian language.

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

S. V. Kasmir Raja, V. Rajitha, and Meenakshi 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.

Sai Rajeswar, Sandeep Subramanian, Francis Dutil, Christopher Pal, and Aaron Courville. 2017. Adversarial generation of natural language.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. pages 273–280.

Max Silberztein. 2016. *Formalizing Natural Languages: The NooJ Approach*.

Alexey Sorokin. 2016. Using longest common subsequence and character models to predict word forms. pages 54–61.

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

Van-Khanh Tran and Le-Minh Nguyen. 2017. Neural-based natural language generation in dialogue using rnn encoder-decoder with semantic aggregation.

Foust William D. 1960. National symposium on machine translation. In *Automatic English inflection*, pages 229–233. UCLA.

Foust William D. 2001. An algorithmic approach to english pluralization. In *Second Annual Perl Conference*. COPE.

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