



<https://doi.org/10.22363/2687-0088-30122>


Research article

A cognitive linguistic approach to analysis and correction of orthographic errors

Robert REYNOLDS^{1,2}, Laura JANDA¹ and Tore NESSET¹

¹*UiT – The Arctic University of Norway, Tromsø, Norway*

²*Brigham Young University, Provo, Utah, USA*

 robert_reynolds@byu.edu

Abstract.

In this paper, we apply usage-based linguistic analysis to systematize the inventory of orthographic errors observed in the writing of non-native users of Russian. The data comes from a longitudinal corpus (560K tokens) of non-native academic writing. Traditional spellcheckers mark errors and suggest corrections, but do not attempt to model *why* errors are made. Our approach makes it possible to recognize not only the errors themselves, but also the conceptual causes of these errors, which lie in misunderstandings of Russian phonotactics and morphophonology and how they are represented by orthographic conventions. With this linguistically-based system in place, we can propose targeted grammar explanations that improve users' command of Russian morphophonology rather than merely correcting errors. Based on errors attested in the non-native academic writing corpus, we introduce a taxonomy of errors, organized by pedagogical domains. Then, on the basis of this taxonomy, we create a set of mal-rules to expand an existing finite-state analyzer of Russian. The resulting morphological analyzer tags wordforms that fit our taxonomy with specific error tags. For each error tag, we also develop an accompanying grammar explanation to help users understand why and how to correct the diagnosed errors. Using our augmented analyzer, we build a webapp to allow users to type or paste a text and receive detailed feedback and correction on common Russian morphophonological and orthographic errors.

Keywords: *morphophonology, phonotactics, orthography, corpus, error taxonomy, webapp*

For citation:

Reynolds, Robert, Laura Janda & Tore Nessel. 2022. A cognitive linguistic approach to analysis and correction of orthographic errors. *Russian Journal of Linguistics* 26 (2). 000–000. <https://doi.org/10.22363/2687-0088-30122>




Когнитивно-лингвистический подход к классификации и исправлению орфографических ошибок

Роберт РЕЙНОЛЬДС^{1,2}  , Лора ЯНДА¹ , Торе НЕССЕТ¹ 

¹Университет Тромсё — Арктический университет Норвегии, Тромсё, Норвегия

²Университет Бригама Янга, Прово, Юта, США

 robert_reynolds@byu.edu

Аннотация.

В представленной статье мы предлагаем систематизацию орфографических ошибок неносителей русского языка на основе лингвистических и когнитивных критериев. Материалом исследования послужили данные лонгитюдного корпуса работ (560К слов) на русском языке, написанных студентами-иностранцами. Традиционные автоматические средства проверки орфографии (спеллчекеры) выявляют ошибки и предлагают исправления, но не могут построить объяснительные когнитивные модели. Предлагаемый подход позволяет распознать не только сами ошибки, но и концептуальные причины этих ошибок, заключающиеся в непонимании фонотактики и морфофонологии русского языка, а также в способах их репрезентации орфографическими правилами. Этот способ позволяет обосновывать причины грамматических ошибок и рекомендовать правила, которые улучшают владение пользователями русской морфофонологией, а не просто исправляют ошибки. Принцип систематизации аннотированных ошибок в корпусе академического письма на неродном языке и таксономия ошибок ориентированы на преподавание. На основе этой таксономии мы разработали набор правил, расширяющих функционал конечно-автоматного анализатора русского языка. Разработанный морфологический анализатор аннотирует словоформы специальными тегами ошибок. Для каждого тега ошибки мы предлагаем сопровождающее пояснение, чтобы помочь пользователям понять, почему и как исправить диагностированные ошибки. Используя наш расширенный анализатор, мы создаем веб-приложение, позволяющее пользователям набирать или вставлять текст и получать подробные комментарии и исправления распространенных морфофонологических и орфографических ошибок в русском языке.

Ключевые слова: морфофонология, фонотактика, орфография, корпус, таксономия ошибок

Для цитирования:

Reynolds R., Janda L., Nessel T A cognitive linguistic approach to analysis and correction of orthographic errors. *Russian Journal of Linguistics*. 2022. Vol. 26. № 2. P. 000–000. <https://doi.org/10.22363/2687-0088-30122>

1. Introduction

Traditional approaches to spell checking are sometimes inadequate for the needs of non-native users because they are optimized for native speakers. Not only is it assumed that the user is capable of choosing between suggested corrections, but the suggestions themselves are optimized for the kinds of errors that *native* speakers make. Even if a non-native user were able to select the correct form from the suggested corrections, it is entirely possible that the user would not understand *why* it is the correct form in contrast to the form they wrote. Furthermore, whereas spell checking for native speakers is mainly a matter of fixing one-off random errors, non-native users need to acquire rules that they can apply in the future. The

mistakes that non-native writers make tend to be systematic, and thereby can be analyzed linguistically and present excellent targeted learning opportunities.

The output of a spellchecker will frequently be either too broad (merely marking a word as misspelled) or too specific (suggesting an alternative for a single given misspelled word) to support the acquisition of useful generalizations. Our proposed tool, the Russian Mentor for Orthographic Rules (RuMOR) is designed to help non-native users connect each specific error to linguistic generalizations, orthographic rules, and examples. This design encourages the user to update their understanding of Russian linguistic and orthographic patterns so that they can avoid making similar errors in the future.

Section 2 reviews related research in the fields of morphological analysis, spelling correction, and intelligent tutoring systems. In Section 3, we describe our methodology, including the process of classifying errors in the RULEC (ETA, April 16, 2022)¹ corpus, modeling the errors in a finite-state framework using *mal*-rules, evaluating the model, and applying the model in a webapp for users. Section 4 contains a summary of our results and future research directions.

2. Related work

Our project is connected to research in a number of disparate fields, including Natural Language Processing (NLP), Intelligent Computer-Assisted Language Learning (ICALL), Russian Linguistics, and Second Language Acquisition (SLA).

2.1. Pedagogical foundations

Textbooks of Russian typically state spelling rules and contain explanations about pronunciation. However, the connection between this material and what it means for confident writing skills is underrepresented. In other words, students may learn that they should pronounce the letter *e* like an *и* when unstressed, or that the letter *u* sounds like *и* when preceded by *и* or *ж*. But students are not warned that these conventions will present challenges in spelling. Furthermore, these rules are typically not exercised in any systematic way and tend to remain peripheral from the students' perspective.

Traditional textbooks take an instruction-based perspective, with the idea of mere transfer of knowledge. A better model for pedagogy is learning by doing, whereby each student constructs their own knowledge network through active engagement. This framework, which is known as constructivism (Biggs 1999, Biggs & Tang 2011), promotes student-centered learning activities both within and outside the classroom. When a student of Russian makes a spelling error, RuMOR can capitalize on that event as an opportunity to engage students with targeted feedback on the relevant spelling and pronunciation conventions. A spelling error is something that is directly relevant to the student in the moment, thus opening up a “teachable moment”, when the student is receptive to improvement of their skills.

¹URL: <http://www.web-corpora.net/RLC/rulec>

When used over time, RuMOR will engage each student with all of the typical errors that they need to focus on.

2.2. Morphological analysis

The Russian language has widespread fusional morphology, with each major word class having multiple inflection classes. Since the complexity of the morphological system is itself the source of many errors, a morphological analysis is frequently essential for determining what feedback will be most helpful to the user. Table 1 shows two authentic orthographic errors which, at the surface level, appear to be the same — mistakenly replacing *u* with *e* — but which are motivated by entirely different parts of the linguistic system.

Table 1. Different underlying motivations for identical surface substitutions

Correct form	Erroneous form	Substitution	Motivation
<i>Марии</i> ‘Maria’	<i>Марие</i>	$u \rightarrow e$	inflectional
<i>умирает</i> ‘dies’	<i>умерает</i>	$u \rightarrow e$	phonological

The erroneous *Марие* is morphologically motivated by the fact that the default Locative singular (and for feminine nouns like this one, Dative singular) ending is *-e*, but the writer has failed to take into account the exceptional rule that nouns in *-ия* take instead the ending *-u*. The incorrect spelling of *умерает* is phonologically motivated by the fact that the pronunciation of *e* is indistinguishable from that of *u* in unstressed syllables, and in all forms of this verb the stress is on the vowel *a*.

The output of traditional spellcheckers would be able to tell the user what substitution is needed to correct the error, but it would be inadequate for determining feedback that helps get at the root of the mistake. On the other hand, a morphological analyzer that is sensitive to the grammatical structure of words can model errors such that these two errors can be linked to distinct and appropriate feedback that is relevant to the different factors that led to the error.

Approaches to automatic morphological analysis of Russian have historically gravitated toward rule- and lexicon-based methods. One reason for this is the existence of the seemingly prescient Grammatical dictionary of Russian (Zaliznjak 1977), which specifies the inflectional patterns of more than 100 000 words. On the basis of this dictionary, computational linguists have produced many Russian morphological analyzers/taggers. These include RUSTWOL (2005) (ENA, April 17, 2022)², StarLing (ENA, April 17, 2022)³ (Krylov & Starostin 2003), DiaLing (ENA, April 17, 2022)⁴, Mystem (ENA, April 17, 2022)⁵ (Segalovich 2003),

² <http://starling.rinet.ru/downl.php>

³ Nozhov, Igor. 2003. *Morphological and Syntactic Text Processing (models and programs)* also published as *Realization of automatic syntactic segmentation of the Russian sentence*. Ph.D. thesis, Russian State University for the Humanities, Moscow. <http://www.aot.ru> (In Russ.)

⁴ <https://yandex.ru/dev/mystem/>; Vilki, Liisa. 1997. RUSTWOL: *A system for automatic recognition of Russian words*. Technical report, Lingsoft, Inc.

pymorphy2 (ENA, April 17, 2022)⁶ (Korobov 2015, Boxarov et al. 2013), and UDAR (ENA, April 17, 2022)⁷. Although all of these analyzers could theoretically be augmented or adapted to provide more informative feedback than a traditional spellchecker, UDAR is best suited to our needs for a number of reasons. First, it is free and open-source, which facilitates operating in an Open Research paradigm. Second, it includes specification of word stress position, which is crucial for predicting some kinds of spelling errors. Third, it is integrated with a Constraint Grammar, a framework designed to deal with inherent ambiguity, a property which errors are notorious for. Fourth, the finite-state paradigm enables extremely fast lookup times, avoiding procedural logic at runtime.

2.3. Spelling and grammar correction

Rozovskaya and Roth (2019) classified errors from the RULEC corpus (The Russian Learner Corpus of Academic Writing, Alsufieva et al. 2012), and found that spelling errors were by far the most frequent class of errors, accounting for 18.6% of non-native errors and 42.4% of heritage speaker errors. Since spelling errors are by definition limited to the modality of writing, it seems safe to say that most, if not all, of these errors are a direct reflection of *writing* proficiency, as opposed to general language proficiency. Therefore, significant improvement in spelling ability is one of the most straightforward paths to build writing confidence and proficiency.

In recent years, there has been a significant uptick in research on spelling correction for Russian (Sorokin 2017), including SpellRuEval, a competition on automatic spelling correction for Russian (Sorokin et al. 2016). However, so far these research projects have understandably been focused only on surface-level correction, without regard to the underlying linguistic sources of the errors. A natural result of this narrow focus is that grammatical input is generally not included because it is not helpful to these models. Whereas grammatical awareness is a sometimes crucial element of pedagogically oriented spelling correction, the official report from SpellRuEval states that adding morphological and semantic features to these models for traditional spelling correction yields little to no gains.

Research on automatic grammatical error correction has been dominated by studies of English, but Rozovskaya and Roth (2019, 2021) have recently extended this research to Russian as well, with impressive results for certain kinds of errors. Although their research path is promising, it falls short for our application in the same way that recent spelling correction does: the training data — and by extension the outputs of the models — do not contain hypotheses about *why* errors are made.

⁶ <https://github.com/kmike/pymorphy2> (based on <https://www.opencorpora.org>)

⁷ <https://github.com/giellalt/lang-rus> and <https://github.com/reynoldsnlp/udar>; UDAR is an abbreviated form of udarénie ‘word stress’, and it is also a recursive acronym: “UDAR Does Accented Russian.”

2.4. Intelligent Language Tutoring Systems (ILTS)

Intelligent Language Tutoring Systems (ILTS) use Natural Language Processing to provide individualized feedback to users without the need for human graders or tutors. Historically, research on ILTS has been focused on workbook-style exercises with tightly controlled context (Heift 2010, Nagata 2009, Amaral & Meurers 2011, Choi 2016; Meurers et al. 2019). In these systems, limiting the context allows the designers to anticipate what kinds of feedback are appropriate. The more controlled the context, the less sophisticated the language analysis needs to be. Conversely, providing feedback on every aspect of language with unlimited context in an ILTS would require something near artificial general intelligence.

One departure from the strategy of tightly controlling the context for feedback in ILTS is the Revita system (Kopotev et al. 2019), which allows users to upload their own texts in a number of languages, including Russian, and generate workbook exercises for that text. Notably, the feedback for incorrect responses is generally limited to connecting the mistake to another word in the sentence that governs the target word, or with which the target word should agree. Unlimited possibilities require limited feedback.

While the goal of RuMOR is also to provide feedback to any arbitrary text entered by the user, it is limited to spelling errors, which tend to be interpretable without reference to any surrounding context. Because the scope of the task is limited to only spelling errors, it is possible to provide detailed feedback with high confidence that the feedback will be germane.

Given the fact that all major Russian morphological analyzers are lexicon- and rule-based, the most natural approach to analyzing Russian produced by non-native speakers in an ILTS is through the use of mal-rules (cf. Sleeman 1982, Mathews, 1992). Mal-rules are rules that are added to license structures that are not valid in the standard language, but are expected in non-native language production. For example, UDAR uses two-level orthographic and phonological rules⁸ to generate standard Russian surface forms from an underlying representation. By modifying or deleting subsets of these rules, one can compile an analyzer that recognizes erroneous wordforms of the sort that non-native writers produce.

3. Methodology

In this section, we describe the methods used to 1) identify the classes of errors to model in our analyzer, 2) augment UDAR to label these errors, and 3) implement the analyzer in the RuMOR webapp.

3.1. Classifying RULEC errors

Russian morphology is more complex than that of many major world languages, and the size of the paradigms, as well as the large number of arcane

⁸ Cf. Koskeniemi, Kimmo. 1983. *Two-level morphology: A general computational model for word-form recognition and production*. Technical report, University of Helsinki, Department of General Linguistics.

exceptions, pose a significant challenge. Although RuMOR is not designed to teach inflectional morphology, there are a number of morphophonological phenomena, such as stem alternations, that directly lead to spelling mistakes. Orthographies tend to accrete idiosyncratic conventions that can be especially obscure to non-native writers, and Russian orthography is rife with challenges. Russian orthography can be characterized as morphophonemic, as it does not always reflect phonological phenomena, such as vowel reduction, consonant voicing assimilation and final-obstruent devoicing.

In order to determine which errors should be included in our model, we turned to the Russian Learner Corpus of Academic Writing (RULEC) (Alsufieva et al. 2012), currently the largest freely available corpus of Russian writing produced by non-native users. It consists of approximately 560K words, written by 15 non-native and 13 heritage writers, all residing in the United States. We analyzed the corpus using the `udar` (ENA, April 17, 2022)⁹ python package to output a list of all words not recognized by the analyzer. This method admittedly overlooks real-word errors, but we suspect that such errors are extremely infrequent in this corpus because opportunities for homophone errors in Russian are mostly limited to a few rare word pairs that are confusable due to final devoicing/voicing assimilation, such as *лук* ‘onion’ vs. *луг* ‘meadow’, both pronounced with final [k].

After generating this list of unrecognized tokens, we constructed a frequency distribution of errors and manually classified the tokens according to whether we believed the token was an actual error, or simply a valid token that UDAR did not recognize, such as the acronym *СПбГУ* ‘Saint Petersburg State University’. For those tokens that we believe are spelling errors, we classified them linguistically according to the motivation behind the error, relying on our expertise as professional linguists and teachers. Each of these error tags is discussed in the following subsections.

The goal of RuMOR is to improve mastery of Russian orthography by making generalizations that users can apply in the future. In this sense, RuMOR has a different and more advanced linguistic goal than that of a spell-checker. Since RuMOR relies on linguistic analysis, it seizes upon spelling errors as teachable moments when it is most appropriate to deliver systematic explanations. Therefore, the tags are linguistically motivated rather than aimed at simple correction. Each tag can be considered an index to link the error to a relevant mini-lesson to help correct the error.

3.1.1. Overview of error tags

Table 2 contains a summary of the error tags currently included in our spelling model and webapp. The “Tag” column is the name of the tag, as implemented in UDAR. Many of the tag names merely describe the substitution that caused the error, so “a2o” means that the letter “a” was erroneously spelled as an “o”. The

⁹ <https://github.com/reynoldsnlp/udar>

“Linguistic label” column is a short pithy description of how to fix the error. More detailed descriptions of the error types are given in the “Tag explanation” column, and relevant examples of misspelled words are provided in the “Examples” column.

Table 2. Summary of error tags

Tag	Linguistic label	Tag explanation	Example(s)
a2o	<i>o</i> → <i>a</i>	Misspelling (<i>o</i> should be <i>a</i>)	<i>озночает</i>
e2je	<i>e</i> → <i>э</i>	Misspelling (<i>e</i> should be <i>э</i>)	<i>ето</i>
FV	no fill vowel	Presence of unnecessary fleeting vowel	<i>отеца</i>
H2S	<i>ь</i> → <i>б</i>	Misspelling (<i>ь</i> should be <i>б</i>)	<i>подъезд</i>
i2j	<i>й</i> → <i>и</i>	Misspelling (<i>й</i> should be <i>и</i>)	<i>миллярд</i>
i2y	<i>ы</i> → <i>и</i>	Misspelling (<i>ы</i> should be <i>и</i>)	<i>близко</i>
ii	<i>ue</i> → <i>и</i>	<i>ue</i> should be <i>и</i>	<i>Марие</i>
lkn	<i>u</i> → <i>е/я/а</i>	Ikanje (<i>u</i> should be <i>е/я/а</i>)	<i>дтей</i>
j2i	<i>u</i> → <i>й</i>	Misspelling (<i>u</i> should be <i>й</i>)	<i>работчи</i>
je2e	<i>э</i> → <i>е</i>	Misspelling (<i>э</i> should be <i>е</i>)	<i>прэкта</i>
NoFV	add fill vowel	Missing fleeting vowel	<i>окн</i>
NoGem	add double letter	Geminate letter is missing	<i>имено</i>
NoSS	add <i>ь</i>	Misspelling (<i>ь</i> is missing)	<i>болше</i>
o2a	<i>a</i> → <i>o</i>	Akanje (<i>a</i> should be <i>o</i>)	<i>каторый</i>
Pal	add softening	Missing palatalization at stem-ending interface	<i>землу</i>
sh2shch	<i>щ</i> → <i>ш</i>	Misspelling (<i>щ</i> should be <i>ш</i>)	<i>лучше</i>
shch2sh	<i>ш</i> → <i>щ</i>	Misspelling (<i>ш</i> should be <i>щ</i>)	<i>вообще</i>
ski	<i>ский</i> → <i>ски</i>	<i>по-ский</i> instead of <i>по-ски</i>	<i>по-русский</i>
SRO	<i>o</i> → <i>е</i>	Spelling Rule <i>o>e</i>	<i>нашой</i>
SRy	<i>ы</i> → <i>и</i>	Spelling Rule <i>ы>и</i>	<i>книгы</i>
y2i	<i>и</i> → <i>ы</i>	Misspelling (<i>и</i> should be <i>ы</i>)	<i>описывают</i>
prijti	<i>прийти</i>	Misspelling the stem of <i>прийти</i>	<i>прийду</i>
revlkn	<i>е/я/а</i> → <i>и</i>	Reversed Ikanje (<i>е, а, я</i> should be <i>и</i>)	<i>умерает</i>
Gem	no double letter	Should be just single, not geminate letter	<i>расширить</i>

3.1.2. Fill vowels

Fill vowels (also known as “fleeting” or “mobile” vowels) are vowels that are only realized if there is no inflectional ending, or if the inflectional ending does not begin with a vowel. For example, *окно* ‘window.SG.NOM’ has an inflectional ending, so there is no fill vowel, but *окон* ‘window.PL.GEN’ has no inflectional ending so the fill vowel appears between the *к* and the *н*.

Fill vowel errors clearly demonstrate both the linguistic motivation for our project, as well as the methodological necessity of a morphological analyzer. There are generalizations that help predict which fill vowels appear in what contexts, but ultimately, they are lexically specified and must be memorized. A traditional spellchecker cannot identify that a particular letter omission or insertion is related to fill vowels, so it cannot direct users to remedial resources. Further, because the “rules” for fill vowels have many exceptions, it is essential to rely on a structured lexicon, such as that in UDAR, to model which errors are related to fill vowels.

We currently have two fill vowel (FV) error tags. The FV tag indicates the presence of a fill vowel that should not be present, and the NoFV tag indicates the absence of a fill vowel that should be present. Since users tend to think in terms of generating oblique forms from the lemma, these tags are far more likely to appear on oblique forms (e.g., erroneous *отеца* ‘father.SG.GEN.FV’ which should be *отца*, and erroneous *окн* ‘window.PL.GEN.NoFV’ which should be *окон*, etc.) as opposed to the lemma, which users are most familiar with, (e.g., errors such as *оту* ‘father.SG.NOM.NoFV’ instead of correct *отец* or *окно* ‘window.SG.NOM.FV’ instead of correct *окно* are quite rare). Our analyzer recognizes all of these forms.

3.1.3. Vowel reduction

Russian vowels are always spelled as they would be pronounced if they were stressed, despite the fact that the sounds of some vowels are very different when they are not stressed. What sounds like unstressed [i] might be spelled *u*, *e*, *a*, or *я*; and what sounds like unstressed [a] might be spelled *a* or *o*. Spelling unstressed vowels is therefore a major challenge, even for native Russian speakers. Native speakers can often solve this problem by remembering a related word or wordform where the given vowel is stressed. For example, to spell *река́* [rik'a] ‘river.SG.NOM’ a native speaker can think of a form of the word with different stress, such as *ре́ку* [r'eku] ‘river.SG.ACC’. However, non-native users have more limited relevant knowledge to draw on, and vowel reduction is one of the most frequent causes for spelling errors in the RULEC corpus.

The pronunciation of an orthographic *o* as [a] is called “akanje” by linguists, and the associated spelling error is tagged *o2a*. The pronunciation of orthographic *e*, *a*, or *я* after palatalized consonants as [i] is called “ikanje”, and the associated spelling error is tagged “*lkn*”. These are the most common error tags for vowel reduction. However, we were surprised to find that *akanje* and *ikanje* create enough confusion in the minds of users that they sometimes do the exact opposite (hypercorrection). The tag *a2o* identifies instances where an orthographic *a* is replaced by *o*, even though it is pronounced [a], as with the token *означае́т* ‘signify.PRS.3P.SG.a2o’ (cf. correct *означае́т*). Similarly, the tag *revlkn* identifies instances where an orthographic *u* is replaced by *a*, *e*, or *я*, as with the token *умирае́т* ‘die.PRS.3P.SG.revlkn’ (cf. correct *умирае́т*).

3.1.4. Phonetic competence

Depending on a user's first language, some of the sounds of Russian are difficult to distinguish, so choosing between letters whose sounds seem indistinguishable is a common problem.

The first instance of confusion that we model is between the letters *ш* and *щ*, both representing voiceless fricatives that English-speaking users associate with “sh”. The prior is post-alveolar, and the latter is palatal. Whether because of the similarity of the orthographic symbols or the similarity of the sounds, non-native writers frequently substitute these letters for one another. The tag *sh2shch* identifies

instances where *u* has been replaced by *u*, as with the erroneous token *лучше* ‘better.ADV.sh2shch’ (cf. correct *лучше*). Conversely, the tag shch2sh marks instances where *u* has been replaced by *u*, as in erroneous *вообще* ‘generally.ADV.shch2sh’ (cf. correct *вообще*).

Another phonetic difficulty is the distinction between the high central unrounded vowel [ɨ] and the high front vowel [i]. Although linguists do not agree on the phonemic status of [ɨ] and [i], they are represented in standard orthography by two separate letters, *ы* and *и*, respectively. Not only is the vowel [ɨ] difficult to pronounce for many non-native speakers, but it is not represented consistently in standard orthography. Although the vowel [ɨ] is mostly represented by the letter *ы*, in some contexts it is written as *и*, most notably when preceded by the letters *ж* or *ш*. The difficulty of phonetic competence, combined with orthographic inconsistency of [ɨ], leads to many spelling errors substituting these letters for one another. The tag y2i marks tokens where *ы* has been replaced by *и*, as in *описывают* ‘describe.PRS.3P.PL.y2i’ (cf. correct *описывают*). The i2y tag marks tokens with the inverse substitution, such as *близко* ‘close.ADV.i2y’ (cf. correct *близко*).¹⁰

Two of our error tags are motivated by a misunderstanding of phonemic palatalization in Russian consonants. In modern usage, the soft sign *ь* indicates that the preceding consonant is palatalized, and the hard sign *ъ* indicates that the preceding consonant is not palatalized. Generally speaking, consonants are assumed to be hard, so the hard sign appears in only one context: between prefixes that end in a consonant, and stems that begin with *е*, *ё*, *ю*, or *я*, as in *подъезд* ‘stairwell’. However, given the relative frequency of the visually similar soft sign *ь*, non-native writers frequently use the soft sign in place of the hard sign, as in *подъезд* ‘stairwell.H2S’ (cf. correct *подъезд*). Similarly, for users that have not acquired palatalization in their language, the role of the soft sign *ь* is difficult to grasp. This leads to its frequent omission, as in *больше* ‘bigger/more.NoSS’ (cf. correct *больше*).

A prominent feature of Russian phonology is consonant palatalization (commonly referred to as hardness vs. softness). Russian orthography marks consonant hardness or softness by two parallel sets of vowel letters (and the symbols *ь* and *ъ*), so that hard consonants are followed by one set, and soft consonants by the other. When inflecting words, users are prone to change the hardness or softness of the stem-final consonant by using a vowel from the wrong set. In particular, it is most common to change soft consonants to hard consonants. Errors of this type are indicated with the tag Pal, as in the error *землю* ‘earth.ACC.Pal’ (cf. correct *землю*).

3.1.5. Alphabetic confusion

Some spelling errors are either evidence of misunderstanding of the sounds or roles associated with a given letter, or interference from the alphabet of the user’s

¹⁰ Note that the i2y tag and the SRy tag are complementary. The i2y tag applies anywhere that the SRy tag does not.

first language. These errors differ from those in Section 3.1.4 (Phonetic competence) in that the users are proficient at producing and perceiving these sounds, but simply fail to associate the sounds with their corresponding symbols. The first pair of such letters is the vowel letter *и* [i] and the consonant letter *й* [j]. Examples of these errors include *работии* ‘worker.SG.NOM.j2i’ (cf. correct *работий*) and *миллярд* ‘billion.SG.NOM.i2j’ (cf. correct *миллиард*).

Another pair of letters that are easily confused are *е* [je] and *э* [e]. The letter *э* only occurs in a small number of high-frequency types, almost exclusively word-initially. Examples of these errors include *это* ‘this.e2je’ (cf. correct *это*) and *проекта* ‘project.SG.GEN.je2e’ (cf. correct *проекта*).

3.1.6. Spelling Rules

A small set of consonant letters have restrictions on which vowel letters are allowed to follow them, in some cases motivated by phonological restrictions at the time of orthographic standardization. The relevant consonants are the so-called hushers (*ж*, *ч*, *ш*, and *щ*), velars (*г*, *к*, and *х*), and the letter *ц*. These spelling rules are generally mentioned by Russian textbooks because they are especially relevant for inflectional endings. However, in many cases textbooks merely state these rules rather than attempting to actively engage students in acquiring them. As a result, such rules tend to remain abstract and students get little opportunity to work out their implications.

The first spelling rule is that after the so-called hushers and *ц*, an unstressed letter *о* is replaced by the letter *е*. Violations of this rule are indicated with the tag SRo, as in the error *нашой* ‘our.FEM.SG.GEN.SRo’ (cf. correct *нашей*).

Another spelling restriction is that after velars or hushers, the letter *ы* is replaced by *и*. Unfortunately, for two of the hushers, this restriction is no longer a valid reflection of modern phonology, since *ж* and *ш* are now non-palatalized consonants. Because of this, not only is the rule sometimes difficult to remember and apply, but it is also phonetically misleading. Violations of this spelling rule are indicated with the tag SRY, as in the error *душы* ‘soul.PL.NOM.SRY’ (cf. correct *души*).

The third spelling rule is one that is not explicitly discussed in any textbooks that we are aware of but is nonetheless a cause for confusion for many non-native speakers. The letter *ц* can be followed by either *ы* or *и*, depending on whether it is in the stem or the inflectional ending. In stems, *ц* is followed by *и* (e.g., *цирк* ‘circus’),¹¹ and in endings *ц* is followed by *ы*. Violations of this rule are indicated with the tag SRc, as in the error *цифровой* ‘digital.SRc’ (cf. correct *цифровой*).

3.1.7. По-__ски

Many adjectives ending in *-ский* can be converted to adverbs by adding the hyphenated prefix “*но-*” and removing the final *й*. For example, *русский* ‘Russian’

¹¹ There are a handful of exceptions to this rule, including *цыплёнок* ‘chick’, *цыган* ‘gypsy’, *на цыпочках* ‘on tiptoe’.

becomes *но-русски* ‘in Russian’. Non-native writers frequently forget to remove the final *й*. This error is indicated by the tag *ski*, as in the error *но-русский* ‘Russian.ski’.

3.1.8. Morphological errors

Another common error is particular to stems ending in an underlying /ij/, whose lemmas orthographically end in *-ий*, *-ие*, and *-ия*, such as *критерий* ‘criterion’, *здание* ‘building’, and *Мария* ‘Maria’. For such stems, any paradigmatic cell that would otherwise end in *-e* ends in *-u* instead. For all three classes, this includes the locative (i.e. prepositional) case and for feminine nouns, the dative case. Errors regarding this principle are indicated with the tag *ii*, as in *о критерие* ‘about the criterion.LOC.ii’ (cf. correct *о критерии*).

3.1.9. Geminates

As in many languages, it is difficult for writers to know which letters are duplicated. Errors that include geminate letters where they should not be are indicated using the tag *Gem*, as in *количество* ‘quantity.Gem’ (cf. correct *количество*).¹² Errors that do not include geminate letters where they should be are indicated with the tag *NoGem*, as in *искусство* ‘art.NoGem’ (cf. correct *искусство*).

3.1.10. Прийти

The stem of the lexeme *прийти* ‘to come’ causes problems for native and non-native speakers alike. The *й* appears in the infinitive *прийти*, but not the indicative: *пришла* ‘come.PST.FEM’, *придет* ‘come.NONPST.3P.SG’. This may feel unexpected when compared with some other prefixed forms of *идти* ‘go’ which do have *й* in the non-past: *зайдет* ‘drop by.NONPST.3P.SG’, *пройдет* ‘pass.NONPST.3P.SG’. Errors related to this lexeme are indicated with the tag *prijti*, as in *прийду* ‘come.NONPST.1P.SG.prijti’ (cf. correct *приду*).

3.2. Automatic error diagnosis: extending UDAR

Each of the sources of errors discussed in Section 3.1 can be formalized in rules defining each of the error types discussed in the previous section. As mentioned in section 2.4, rules that license non-normative words or structures are referred to as mal-rules (cf., e.g., Sleeman, 1982; Matthews, 1992, and references therein). In this section, we provide an abbreviated overview of the mechanics of applying our mal-rules to UDAR.

¹² The insertion of geminates is problematic for practical reasons. The corresponding mal-rule would apply to virtually every letter of every word in the analyzer, exploding the amount of storage/memory required for the analyzer. Although theoretically possible, the *Gem* tag is usually omitted for practical reasons.

UDAR is a finite-state transducer, built using three formalisms: the *lexc* language for creating the finite-state lexical network; the *twolc* language for realizing orthographic and morphophonological rules on surface forms; and *vislcg3* for writing a Constraint Grammar to resolve morphosyntactic ambiguity on the basis of surrounding context.¹³ Our mal-rules are applied in one of two ways. First, rules that are sensitive to underlying morphophonological structure—such as *ii*, *FV*, *NoFV*, *Pal*, and *SRO*—are implemented as alternative *twolc* rules.¹⁴ Rules that can be modeled as simple character substitution are implemented as XFST regular expression replace rules.¹⁵ In either case, the process for adding a tag to the transducer is the following.

First, a standard transducer is compiled, using UDAR’s original rules. Then, for each tag, the mal-rule is applied to make an error transducer. The standard transducer is subtracted from the error transducer so that only wordforms that were affected by the mal-rule remain. Then, the error tag is added to all forms in the error transducer, and the resulting transducer is added to the standard transducer by disjunction. (ENA, April 17, 2022)¹⁶. In this way, all of UDAR’s original contents are preserved, and all additions are tagged with the appropriate error tags.

To the extent possible, errors are accumulated, one after the other, so that words with more than one kind of error can be recognized. However, several of the rules feed into one another, or could even reverse one another. For example, if *e2je* were added on top of *je2e*, the resulting surface form would be identical to the correct form, but would be tagged for both errors. Therefore, the errors were grouped by contexts, and all errors affecting the same context are added in parallel. In this way, errors in different context-groups can stack on one another, but errors in the same context-group do not.

3.2.1. Evaluation

We analyzed the entire RULEC corpus using our augmented analyzer, compiled a list of all types that are tagged as errors, and compared the output of the analyzer with our manual labels. We found that for our target errors, the analyzer has perfect recall, meaning that every token that was manually labelled with one of our target error tags was also labeled by the augmented analyzer as such. However, not all of the errors identified in the corpus fit into these categories. Out of 279 manually labeled error types, our analyzer labeled 124 (44.4%). Out of 999 manually labeled error tokens, our analyzer labeled 467 (46.7%).

¹³ The *lexc* and *twolc* source files can be compiled using either Xerox Finite-State Tools (XFST) (Beesley and Karttunen 2003) or Helsinki Finite-State Transducer Technology (HFST) (Linden et al. 2011).

¹⁴ For a detailed explanation of how the *twolc* rules in UDAR function, see chapter 2 of Reynolds, Robert. 2016. *Russian natural language processing for computer-assisted language learning: capturing the benefits of deep morphological analysis in real-life applications*. Ph.D. thesis, UiT – The Arctic University of Norway.

¹⁵ For a detailed explanation of XFST regular expressions, see Beesley and Karttunen (2003).

¹⁶ The Makefile that builds the error transducer can be found at https://github.com/giellalt/lang-rus/blob/8839887e986ae15a255e3396f08d394e8efac363/src/Makefile_L2

3.3. RuMOR webapp

RuMOR is a free and open-source webapp allowing users to get interactive feedback on Russian spelling errors. (ENA, April 17, 2022)¹⁷ RuMOR was built as a mobile-first webapp, so that it can be used comfortably on desktops, laptops, and mobile devices. Currently, two interface languages are available: English and Norwegian. A screenshot of the app is shown in Figure 1.

The user is prompted to type or paste a text, and upon submitting the text, words identified by our augmented analyzer as spelling errors are turned into clickable links. Tokens are considered errors only if all possible readings are errors, so our system does not currently attempt to handle real-word errors. For example, in Figure 1, the token *эй* ‘hey’ is obviously intended to be *ей* ‘she.DAT’, but because the analyzer outputs at least one non-error reading, it is not treated as an error by RuMOR.¹⁸

When an error is clicked, all possible readings are shown in a pane to the side of the text. For each reading, we display the dictionary form, the type of error that would lead to the attested token, and the corrected form (which is shown by clicking or hovering). The readings are sorted by lemma frequency, so the most likely reading is listed first. In Figure 1, the token *Ана* is selected, and four possible readings are displayed: *она* ‘she.o2a’, *оно* ‘she.o2a’, *Анна* ‘Anna.NoGem’, and *Аня* ‘Anya.Pal’.

When the user clicks on any of the error tags, the error explanation is shown in the next column. These explanations are intended to be as short as possible while still giving enough explanation and examples to be reasonably complete. The explanations are open-source, and hosted separately at (ENA, April 17, 2022)¹⁹

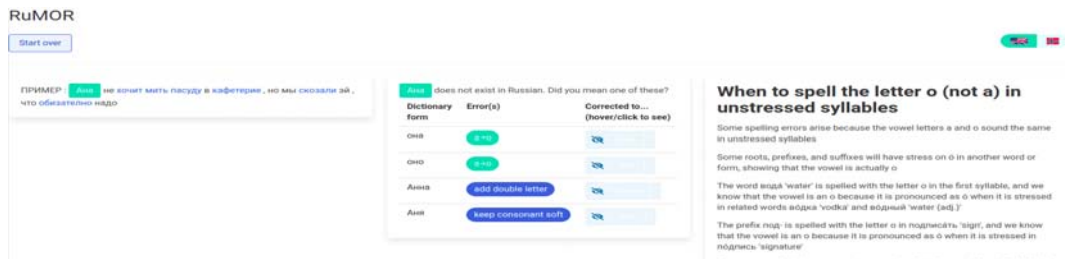


Figure 1. Screenshot of the RuMOR webapp

4. Conclusions and future work

This article has introduced RuMOR, a free, open-source, interactive webapp for identifying, diagnosing, correcting, and explaining a variety of common spelling

¹⁷ The source code for the webapp is available at https://github.com/reynoldsnlp/rus_L2_flask. At the time of writing, the app is accessible at <https://icall.byu.edu/rumor>.

¹⁸ Although this particular example would be difficult to disambiguate, some real-word errors can be resolved by Constraint Grammar rules which would remove some real-word readings on the basis of the surrounding context.

¹⁹ https://github.com/reynoldsnlp/rus_grammar_explanations.

errors, based on linguistic analysis. The webapp uses a modified version of the UDAR analyzer, which we augmented using mal-rules. The validity of our model was maximized by deriving error tags from real-world errors identified in the RULEC corpus. To our knowledge, this is the first such application for Russian that attempts to provide comparable targeted feedback to any arbitrary running text.

This linguistic approach is especially well-suited to error annotation, but also facilitates text normalization. As demonstrated in the webapp, UDAR can automatically generate the corrected wordform.

Another potential application of our error-augmented analyzer is automatic corpus annotation. Until now, corpora of Russian texts produced by non-native speakers have relied almost exclusively on human annotators to analyze and classify errors. Our analyzer can make this process faster and more consistent by giving annotators a preliminary linguistic analysis of orthographic errors to review.

Future work will focus on adding more classes of errors attested in corpora. These errors include conjugation errors, especially related to stem alternations and inflection class selection. Hapaxes in RULEC were excluded from the present study, but we know that there are some error types represented among them that deserve to be included in our error model. For example, users whose first language uses the Latin alphabet frequently misuse alphabetic false friends, i.e., letters that appear the same as Latin letters, but which represent different sounds. In addition to expanding our spelling error model, we also intend to expand UDAR's existing Constraint Grammar to add syntactic error labels.

Finally, although it is tempting to assume that RuMOR is an effective tool, it is crucial to understand how such tools are actually used, and what effect they have on motivation and proficiency outcomes. We hope to perform evaluations and experiments to understand the outcomes of this project.

REFERENCES

- Beesley, Kenneth R. & Lauri Karttunen. 2003. *Finite State Morphology*. Stanford, CA: CSLI Publications.
- Biggs, John. 1999. What the student does: Teaching for enhanced learning. *Higher Education & Development* 18 (1). 57–75.
- Biggs, John & Catherine Tang. 2011. *Teaching for Quality Learning at University*. Maidenhead, UK: Open University Press.
- Boxarov, V.V. et al. 2013. Crowdsourcing morphological annotations. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog"* 1. http://opencorpora.org/doc/articles/2013_Dialog.pdf (accessed 20.04.2022).
- Kopotev, Mixail et al. 2019. Corpus-based language teaching tool. *Trudy Meždunarodnoj Konferencii «KORPUSNAJA LINGVISTIKA–2019»*. 30–39. (In Russ.)
- Korobov, Mikhail. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In *Proceedings of AIST'2015*. 320–332. New York: Springer.
- Krylov, Sergej & Sergej Starostin. 2003. Upcoming tasks for morphological analysis and generation in the integrated information environment STARLING. In *Proceedings*

- of the International Conference “Dialog 2003”. <https://www.dialog-21.ru/media/2655/krylov.pdf> (In Russ.) (accessed 20.04.22).
- Linden, Krister et al. 2011. HFST– framework for compiling and applying morphologies. In Cerstin Mahlow & Michael Pietrowski (eds.), *Systems and frameworks for computational morphology*, 100 of Communications in Computer and Information Science. 67–85. New York: Springer.
- Matthews, Clive. 1992. Going AI: Foundations of ICALL. *Computer Assisted Language Learning* 5(1). 13–31.
- Rozovskaya, Alla & Dan Roth. 2019. Grammar Error Correction in Morphologically Rich Languages: The Case of Russian. *Transactions of the Association for Computational Linguistics* 7. 1–17. https://doi.org/10.1162/tacl_a_00251
- Rozovskaya, Alla & Dan Roth. 2021. How Good (really) are Grammatical Error Correction Systems? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2686–2698.
- Segalovich, Ilya. 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*. 273–280.
- Sleeman, Derek. 1982. Inferring (mal) rules from pupil’s protocols. In *Proceedings of the 5th European Conference on Artificial Intelligence (ECAI)*. 160–164. Orsay, France.
- Vilki, Liisa. 2005. RUSTWOL: A tool for automatic Russian word form recognition. In Antti Arppe et al. (eds.), *Inquiries into words, constraints and contexts: Festschrift for Kimmo Koskenniemi on his 60th Birthday*, 151–162. Stanford, CA: CSLI Publications.

Dictionaries

- Zaliznjak, Andrej A. 1977. Grammatical dictionary of the Russian language: Inflection: Approx 100 000 words. *Russkij Jazyk*. (In Russ.)

Article history:

Received: 20 October 2021

Accepted: 21 January 2022

История статьи:

Bionotes:

Laura A. JANDA is Professor of Russian in the Department of Language and Culture at UiT The Arctic University of Norway. She holds a PhD in Slavic Linguistics from UCLA (1984). She pursues research in the framework of cognitive linguistics applied mostly to the analysis of grammatical categories and constructions in Russian using corpus data. She also works on the development of research-based electronic resources for learners of Russian.

Contact Information:

UiT The Arctic University of Norway

UiT Noregs arktiske universitet

Postboks 6050 Langnes 9037 Tromsø

e-mail: laura.janda@uit.no

ORCID: <https://orcid.org/0000-0001-5047-1909>

Tore NESSET is Professor of Russian linguistics in the Department of Language and Culture at UiT The Arctic University of Norway. He received his doctoral degree from the University of Oslo in 1997. His research interests include corpus and cognitive linguistics which he applies to the study of Russian and Norwegian. He also works on historical linguistics and is the author of the widely used textbook *How Russian Came to Be the Way It Is* (2015).

Contact Information:

UiT The Arctic University of Norway
UiT Noregs arktiske universitet
Postboks 6050 Langnes 9037 Tromsø
e-mail: tore.nesset@uit.no
ORCID: <https://orcid.org/0000-0003-1308-3506>

Robert REYNOLDS is employed as Assistant Research Professor in the Office of Digital Humanities at Brigham Young University. He holds a PhD in Russian Language Technology from UiT The Arctic University of Norway. His research interests include Intelligent Computer-Assisted Language Learning (ICALL); Natural Language Processing for low-resource languages; automatic analysis of text complexity/readability; automatic reading proficiency assessment using eye-tracking; structure of Russian; and morphological complexity.

Contact Information:

Brigham Young University
Brigham Young University Provo, UT 84602
e-mail: robert_reynolds@byu.edu
ORCID: <https://orcid.org/0000-0003-0306-087X>.

Сведения об авторах:

Лора А. Янда – профессор факультета языка и культуры Арктического университета Норвегии, степень доктора наук получила в Калифорнийском университете в Лос-Анджелесе (1984), специалист по славянскому языкознанию. Сфера интересов включает когнитивную и корпусную лингвистику, грамматические категории русского, а также создание электронных ресурсов исследовательского типа для изучающих русский язык.

Контактная информация:

UiT The Arctic University of Norway
UiT Noregs arktiske universitet
Postboks 6050 Langnes 9037 Tromsø
e-mail: laura.janda@uit.no
ORCID: <https://orcid.org/0000-0001-5047-1909>

Туре Нессет – профессор кафедры языка и культуры Арктического университета Норвегии. Докторскую степень получил в Университете Осло в 1997 году. Его исследовательские интересы включают корпусную и когнитивную лингвистику применительно к русскому и норвежскому языкам. Он также работает в области исторической лингвистики и является автором широко известного учебника *How Russian Came to Be the Way It Is* (2015).

Контактная информация:

UiT The Arctic University of Norway

UiT Noregs arktiske universitet

Postboks 6050 Langnes 9037 Tromsø

e-mail: tore.nesset@uit.no

ORCID: <https://orcid.org/0000-0003-1308-3506>

Роберт Рейнольдс работает доцентом-исследователем в отделе цифровых гуманитарных наук Университета Бригама Янга. Имеет докторскую степень по языковым технологиям в русском языке, полученную в Арктическом университете Норвегии. Его исследовательские интересы включают обучение языку с помощью интеллектуальных компьютерных технологий (ICALL); обработку естественного языка для малоресурсных языков; автоматический анализ сложности/читабельности текста; автоматическую оценку навыков чтения с помощью айтрекинга; структуру русского языка и морфологическую сложность языков.

Контактная информация:

Brigham Young University

Brigham Young University Provo, UT 84602

e-mail: robert_reynolds@byu.edu

ORCID: <https://orcid.org/0000-0003-0306-087X>.