NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Nikolay V. Karpov, Vera G. Sibirtseva

TOWARDS AUTOMATIC TEXT ADAPTATION IN RUSSIAN

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS

WP BRP 16/LNG/2014

Nikolay V. Karpov[1], *Vera G. Sibirtseva*[2]

TOWARDS AUTOMATIC TEXT ADAPTATION IN RUSSIAN

This article describes ways of using original texts in the National Russian Corpus and news texts to teach Russian as a foreign language. The two-year work of a scientific group from the Higher School of Economics (from Nizhny Novgorod and Moscow), called CorpLings was analyzed. Special attention was paid to the automatic adaptation of acute news texts, which was the basic principle of the research part of the project. We also describe ways of simplifying syntactical and morphological structures that may seem difficult for students at an elementary level. The stages used for lexical simplification are described in detail, such as the creation of an algorithm to find the most appropriate synonyms based on morphological rules, and an analysis of the statistical model of words' contextual proximity. This article also addresses the difficulties faced by developers and the final results of our research.

Keywords: Russian, electronic textbook, text simplification, contextual proximity, distributional semantic model.

JEL: Z

[1] National Research University Higher School of Economics, Department of Applied Mathematics and Informatics, nkarpov@hse.ru

[2] National Research University Higher School of Economics, Department of Applied Linguistics and Intercultural Communication, vsibirtseva@hse.ru

1. Introduction

Due to globalization, national languages have garnered more interest all over the world, and Russian is no exception. More and more foreign students want to learn the language of Tolstoy and Pushkin. Unfortunately, when students start learning, they often have a lack of manuals or textbooks that could have helped them acquire the language efficiently. The language of news texts and articles is especially difficult to learn.

To fill the gap of information which can be adapted for a student's learning needs, a scientific group of students and professors from the Higher School of Economics launched a project called «CorpLing». The project's main goal is to create helpful electronic multimedia manuals for students who are planning to learn Russian as a foreign language.

The project was implemented in two stages. In the first stage we created an electronic textbook, using the "eFront" learning management system, called «Russian verb». The manual is focused on the productive prefix word-formation of Russian verbs and intended for advanced students. Foreign students were offered to choose any prefix from the list and read details about it and its meaning. The theoretical material was accompanied by examples from modern texts about various topics. Moreover, the stress marks that are provided throughout the tutorial provide optimal training to correct students' intonation and pronunciation.

We used the Russian National Corpus (RNC), because it contains different types of language material. More than 500 million words help us to provide a tutorial from literature, scientific, business and the everyday use of words.

An undisputed advantage of using the RNC to create new interactive textbooks is the ability to determine the most frequent grammatical and lexical constructions. Unlike formal artificial examples, which are widely used in paper-based textbooks, our electronic textbook helps foreigners to focus on everyday language and acquaints them with language changes. The wide coverage of vocabulary associated with fiction, media, online blogs, forums and other areas of communication allows foreigners to expand their vocabulary in the professional sphere.

This stage of the project showed that using the RNC's authentic materials as a basis for the textbooks proved to be effective, but at the same time quite difficult. First of all, Russian vocabulary and syntax are quite difficult to understand. Texts that reflect a living Russian language are usually taken from blogs and forums, where Internet users do not always express their thoughts in correct Russian and often use slang words. Many of the selected phrases often contain specific terms which were particular to a kind of special activity or profession. These words are difficult for foreigners to understand, and so we tried to avoid them when creating a textbook. The editing process was manually implemented and took a considerable amount of time to carry out.

Having discovered the complexity of the manual adaptation process, our group came up with another idea, and we decided to improve the project using an automatic process. We wanted to reduce the number of people spending time on adapting texts and exercises according to a student's language level. Therefore, the second part of the «Russian Verb» project was the development of using adapted journalists' texts. Our group switched from scientific and applied tasks to experimental research in the second stage.

Section 2 focuses on the problem of using adapted texts in teaching and commercial activities. In Section 3, we prototype an algorithm for text and single sentence retrieval with needed ratability level. In Section 4 we empirically explore how people simplify a text. Section 5 proposes an approach for the automatic adaptation of texts to a lower level of readability. In Section 6 we present a case study of text adaptation. Section 7 concludes the paper.

2. Using adapted texts in teaching and commercial activities

Internet users currently generate a large amount of content, in any language of the world, through news, analytical news articles or users' comments. Using these types of texts seems to be useful when learning a foreign language, as reading in a foreign language shapes the perception of new and relevant information in another language. Understanding the content of articles helps students to develop their language skills and educate them about the culture

The main obstacle for a student learning a foreign language through contemporary texts is the complexity of the language used in the news. Authors normally write for native speakers, who can easily overcome language novelties, while for a student learning a language it is important for a text to be adapted to their level. The most difficult things like low-frequency vocabulary and complex syntactic constructions must be changed to simpler and more common ones. If the text is to be brought in line with the student's language level and simplified, significantly revising the text can sometimes be as difficult as writing a new one. The process of rewriting for the student's level is called abridgment or adaptation and is done manually. Adapted or abridged texts are undoubtedly easier to read for someone who is not a native speaker (Chandrasekar et al., 1996).

Text abridgement takes any professional a long time to do, and since it is so time consuming, the process can become quite expensive. It has therefore not been used for the large volumes of information on the internet. However, we were interested in giving this information to our students, and so we started to think about making this process less time consuming and expensive by using computer technology.

There is another reason to use news texts. Search engines rank internet resources by taking into account the speed with which they are updated. News web-sites which create the so-called "second level" news (created from open sources and news agencies), fight for visitors and search engines' high rankings. The task of the person rewriting a text is to write an article about a hot topic, using new expressions and language structures to make the text look unique. A copy writer operates the same facts, writes about the same people and, geographic location, but using different expressions. As a result, the network generates a lot of texts. Therefore, our idea is to use this fact to select synonymous words and expressions from the news.

The second stage of our work's objective was to study approaches that would help us to automate the abridgment process of news texts. This stage began with a deep study of syntactic structures and the lexical complexity of sentences in the existing Russian language textbooks which are available for foreign students. The results show that these sentences are adapted for students with a very limited knowledge of language structures. For example, textbooks do not contain complications such as a recursive chain of subordinate clauses, involving constructions and adverbial participles, information in parentheses or the formal ways of expressing direct speech. Dictionaries and glossaries only include the active lexical minimum, which can correspond to a certain level of competence in the language, excluding very difficult words. At the same time, news texts and examples of the RNC represent the true base of the language and a level of syntactic complexity which is often much higher than in existing textbooks.

We took part of the morphologically marked corpus SynTagRus to sample the basis of the syntactic structures (Nivre et al., 2008). All sentences in SynTagRus were analyzed as "simple", corresponding to the threshold level of proficiency in Russian as a foreign language, and «complicated» respectively.

In this research, the concept of "simple" sentences does not have the same meaning as generally used in linguistics and syntactic field. Based on the requirements to study Russian as a foreign language for A1 and A2 level, we formulate the main characteristics of "simple" sentences for specific academic purposes; a sentence length from 2 to 10 words, no participles

and a gerundial clause; no more than three homogeneous members and compound and complex sentences split into simple ones. Elliptical sentences are not taken into account.

Examples:

И цена на них была высокой.
Одним из основных продуктов питания на побережье была рыба.

During the second stage, we automatically identify the text's readability and convert complex parts into "simple" ones. The final product, which must automatically simplify the text, is expected to be an intermediary between students and the source. Identifying the rules for transforming the syntactic structure without changing the meaning may encourage more theoretical research in the cognitive domain.

Among further steps, we carried out a detailed study of how texts are adapted by linguistic rules and we worked with lexical analyzer and frequency dictionary. We also began to programme add-ons that would allow us to adapt and simplify the language structures and configure and integrate lexical analyzers.

3. Readability prediction

The first step was to apply the models which had been developed to predict readability for children, to readability prediction for Russian as a foreign language. We extracted features from a collection which consisted of 219 texts divided into four groups. The levels were distributed as follows: A1 (elementary) – 52 texts, A2 (basic) – 57, B1 (first) – 60, C2 (difficult) – 50 texts according to the levels in the Common European Framework of Reference for Languages (CEFR) (Verhelst et al., 2009). The first three groups included texts created especially for second language learners of Russian through news articles, bearing in mind their level of language knowledge[3].

Some classical methods for identifying the readability level are well-known. The Dale-Chall model, the Flesch-Kincaid model grade level and Mackovsky's model are widely used, alongside others, in modern online readability predictors[4][5]. The Flesch-Kincaid model [2] discussed the complexity of English texts as a linear function of the average number of syllables per word and the average length of the sentence. An output value relates to a U.S. grade level, or the number of years of education required to understand the text.

$$R_{Fl\text{-}K}=[(11.8\times ASW)+(0.39\times ASL)-15.59] \quad (1)$$

The ASL is the average sentence length (number of words divided by the number of sentences) and the ASW is the average number of syllables per word (number of syllables divided by the number of words).

Dale and Chall's formula [3] also defines the syntactic difficulty of the text as the average length of the sentence, but for the lexical variable it uses the percentage of words which are not from the list of 3000 Simple Words (NSW), which is based on words' familiarity. This means that all the words in the list would be familiar to US children in the 4th grade.

$$R_{D\text{-}Ch}=[(0.1579\times NSW)+(0.0496\times ASL)+3.6335] \quad (2)$$

This formula developed for children who have a low level of language proficiency.

The automatic identification of reading difficulty in Russian is also researched in a number of works. Oborneva's (2006) work adapts Flesch and Flesch-Kincaid's formula for Russian by adjusting the coefficients. She compares the average length of syllables in English

[3] http://texts.cie.ru
[4] http://ru.readability.io/
[5] https://readability-score.com/

and Russian words and the percentage of multi-syllable words in dictionaries for these languages.

$$R^{Ru}_{Fl\text{-}K}=[(8.4\times ASW)+(0.5\times ASL)-15.59] \quad (3)$$

Using her adaptation method to Russian, we change a coefficient in the Dale and Chall formula according to the ratio of the average sentence length in both languages. Most of our foreign students have a low level of Russian language proficiency and we have a list of words that are recommended to be used in the students' active vocabulary. We can calculate the NSW variable as the number of words not included to this list. Therefore, Dale and Chall's readability formula for Russian as a foreign language looks like this:

$$R^{Ru}_{D\text{-}Ch}=[(0.1579\times NSW)+(0.062\times ASL)+3.6335] \quad (4)$$

The number of non simple words (NSW) is counted as the number of words which are not included in the set of our target level's lexical minimum (Andriushina, 2011).

A readability formula was developed especially for Russian children by Mackovskiy (1973). He used the standard regression technique and analyzed 50 texts marked by 60 children and two factors: non simple words and the average sentence length.

$$R^{Ru}_{Mac}=[(0.123\times NSW)+(0.62\times ASL)+0.051] \quad (5)$$

Non simple words (NSW) here are taken to mean any words with more than 3 syllables.

We calculate the empirical distribution functions of $R^{Ru}_{D\text{-}Ch}$, $R^{Ru}_{Fl\text{-}K}$ and R^{Ru}_{Mac} for each text from the 4 readability levels. Using empirical distribution, we estimated the parameters of the Gaussian function for each subset. The results are shown in Figure 1.

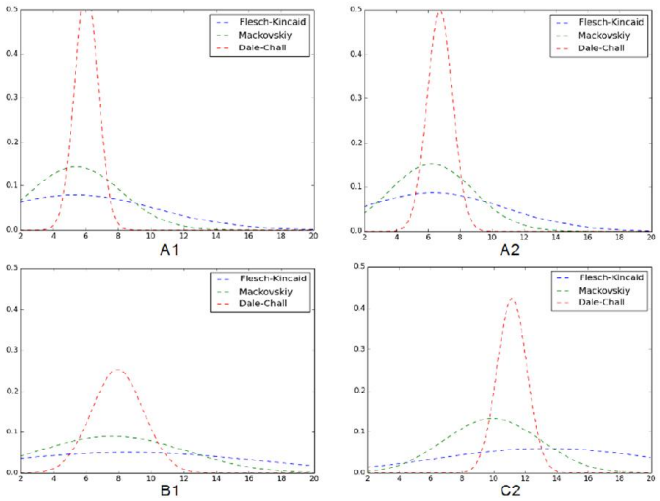

Figure 1 Estimated distribution of values $R^{Ru}_{D\text{-}Ch}$, $R^{Ru}_{Fl\text{-}K}$ and R^{Ru}_{Mac}

Given these results, some mean values of distribution function are comparable but the standard deviations are quite different for different methods. The lowest standard deviations are found in the Dale and Chall formula. The Flesch-Kincaid model shows the highest deviations, and Mackovskiy's model seems to have the best fit.

The second task was to perform the prototyping retrieval of Russian texts with needed readability. The main goal of this process was to discover which variables and classification algorithms would allow us to obtain the highest indicators of precision and recall of readability prediction. The evaluation was performed using cross validation on the test part of our collection with 219 texts. Three features from previous models (ASL, ASW and NSW) we complemented by 22 features proposed in previous work:

1. Average number of words in the sentence of the text

2. Average length of one word in a sentence
3. Text length in letters
4. Text length in words
5. Average sentence length in syllables
6. Average length of words in syllables
7. Percentage of words with number of syllables equal to or more than *N*. We define *N* as a value from 3 to 6
8. Average sentence length in letters
9. Average length of words in letters
10. Percentage of words with number of letters equal to or more than *N*. We define *N* as a value from 5 to 13
11. The percentage of words in a sentence not included in the A1 level's active vocabulary
12. The percentage of words in a sentence not included in the A2 level's active vocabulary
13. The percentage of words in a sentence, not included in the B1 level's active vocabulary
14. The occurrence of concrete parts of speech in the sentence

We marked seventeen parts of speech in the texts according to the list of grammar marks in the OpenCorpora (Bocharov et al., 2011):

- noun (NOUN)
- full form of an adjective (ADJF)
- short form of an adjective (ADJS)
- comparative (COMP)
- personal form of the verb (VERB)
- infinitive form of the verb (INFN)
- full participle (PRTF), short participle (PRTS)
- gerund (GRND)
- numeral (NUMR)
- adverb (ADVB)
- noun-pronoun (NPRO)
- predicative (PRED)
- preposition (PREP)
- conjunction (CONJ)
- a particle (PRCL)
- interjection (INTJ)

We were interested in occurrence of parts of speech as proposed by (Franccois, 2009).

We did not use some of the variables described in (Nevdah, 2008) due to the way that our texts were adapted, nor did we use variables connected with paragraphs, given that our texts are very short. The texts do not have syntactic markup, which is why the concept of a phrase was not used either.

We use a well known classification algorithm (Classification Tree, SVM and Logistic Regression, Random Forest) to conduct a series of experiments about retrieving texts with the necessary level of readability. Due to the fact that the results of the SVM technique reached 98%, it is possible to say that the results meet the needs. This high result could be explained by the fact that the collection of texts used for the evaluation was created especially for foreign students, and seem a bit artificial for native speakers.

The next task was to make a prototype of an algorithm to retrieve difficult single sentences for further simplification. This algorithm is based on a sentence which was classified according to its readability. 25 variables from the texts mentioned above were adapted to be extracted from a single sentence. Traditional classification techniques like the Flesch-Kincaid

and the Dale-Chall model were also adapted to identify the lexical and structural complexity of a single Russian sentence.

We used a subcorpus of the Russian national Corpus (RNC) called the SynTagRus corpus (Nivre et al., 2008) to evaluate our results, which has morphological and syntactic metadata. We manually tagged 3500 sentences from this subcorpus to mark their structural level of perception complexity. We found out that level B1 suits the majority of our students, and so we created a binary sentence markup: 1) B1 or lower than B1 and 2) Higher than B1. The best results for the single sentence readability prediction are shown in Table 1.

Table 1. Single sentence readability prediction using all variables and syntactic links

Method	Classification accuracy	F-measure (difficult /simple)	Precision	Recall
Naive Bayes	0.8191	0.8906/ 0.4767	0.8354/ 0.6975	0.9537/ 0.3621
Knn	0.8224	0.8893/ 0.5501	0.8571/ 0.6493	0.9241/ 0.4772
Random Forest	0.9443	0.9640/ 0.8768	0.9620/ 0.8832	0.9661/ 0.8705
Classification Tree	0.9364	0.9584/ 0.8648	0.9679/ 0.8380	0.9491/ 0.8933
SVM	0.8633	0.9125/ 0.6875	0.9679/ 0.7165	0.9491/ 0.6607

All of the features (including statistical, lexical and syntactical) can predict sentence readability with a 0.9661 recall using the Random Forest algorithm. The most important features for this classification are lexical. To analyze the contribution of a particular variable, we counted information gain value and test accuracy with a different subset of variables.

Table 2. Results of total readability prediction using all variables and syntactic links

Variable name	Information gain ratio
The percentage of words in a sentence are not included in the active vocabulary of the B1 level	0.318
Sentence length in letters	0.122
Percentage of words with 3 syllables or more	0.119
Sentence length in syllables	0.118
Sentence length in words	0.098
Syntactic predicative link	0.095
Average words length in syllables	0.092
The average length of one word in a text	0.092
Percentage of words with 7 letters or more	0.069
Percentage of words with 5 letters or more	0.069

Finally, we found one variant of the model to effectively identify the readability of Russian sentences using syntactic links. The results indicate that knowing the syntactic structure weakly increases accuracy. Structural complexity is highly correlated with sentence length. For a more detailed description see Karpov et al.'s (2014) article.

4. Empirical analysis. Morphological and syntactic adaptation.

To construct the automated system of text adaptation, we empirically explored how people simplify texts. We took a set of news texts with different kinds of topics, including 10 texts from the RIA news agency's web site. We then asked independent experts who specialize in teaching Russian as a foreign language to adapt these texts manually to an elementary level of the language. At the same time, we recorded the methods that the experts used. When they finished, each expert wrote a report, where all the methods used for adapting the text were systematized.

The original text:

Эксперты "ЛК" провели исследование трех российских социальных сетей: "ВКонтакте", "Одноклассники" и "Мой Мир". В ходе недельного исследования эксперты выяснили, что ребенку в возрасте 13 лет в соцсети "ВКонтакте" по запросу "порно" выдается список из закрытых групп, но при введении в поисковый запрос синонимов этого слова ребенок получает перечень открытых групп, содержащих информацию по запрашиваемой тематике, и может без проблем ее просмотреть. Кроме того, ребенок не застрахован от общения с педофилами в соцсети, может подвергаться моральному насилию и запугиванию, а также столкнуться с мошенниками. В социальной сети "Одноклассники" по запросу "порно" и его синонимов ребенку становится доступен список закрытых групп, в которые нельзя получить доступ. В этой социальной сети подросток может рассчитывать только на "легкую эротику". Общение с педофилами в "Одноклассниках" сведено к минимуму, а каких-либо видов мошенничества эксперты в сети не обнаружили.

This is the first adapted variation:

Эксперты "ЛК" изучили три российские социальные сети: "ВКонтакте", "Одноклассники" и "Мой Мир". За неделю исследования эксперты поняли, что в соцсети "Вконтакте" по запросу "порно" ребенок в возрасте 13 лет видит закрытые группы. Но при поиске синонимов этого слова ребенок получает открытые группы. В них есть порно, и ребенок может получить легко его смотреть. Кроме того, ребенок может общаться с педофилами в социальной сети. Он может встретиться с моральным насилием. Также он может встретиться с мошенниками. В социальной сети "Одноклассники" по запросу "порно" и его синонимов ребенок видит закрытые группы, в которые он не может войти. В этой социальной сети подросток может получить только "легкую эротику". В "Одноклассниках" почти нет общения с педофилами. Также эксперты не нашли мошенничества в этой сети.

The second adapted variation:

Эксперты исследовали три российских социальных сети: "ВКонтакте", "Одноклассники" и "Мой Мир". Ребенок 13 лет в соцсети "ВКонтакте" по запросу "порно" находит закрытые группы. При поиске синонимов этого слова ребенок получает открытые группы с информацией по этой теме, и может ее посмотреть. Ребенок может общаться с педофилами в социальной сети. Он может подвергаться моральному насилию и запугиванию, а также встретиться с мошенниками. В социальной сети "Одноклассники" по запросу "порно" и его синонимов ребенок находит закрытые группы. В этой социальной сети подросток может смотреть только "легкую эротику". В "Одноклассниках" почти нет общения с педофилами, и мошенничества эксперты в этой сети не нашли.

Having compared the results, we offered a list of rules that described the way in which sentences are morphologically adapted.

Examples of noun adaptation:

a) the noun/verb (which can be replaced by a nominal predicate) and a noun, formed through a verb nominalizing transformation, are replaced with a nominal predicate + verb:

о ***необходимости принятия*** *мер* **=>** *о том, что* ***необходимо принять*** *меры административного характера*

требует *долгого* ***тестирования и отладки*** **=>** ***нужно*** *много* ***тестировать*** *и* ***отлаживать***

b) abbreviations and acronyms are replaced by full forms of the words or general synonyms

Минобрнауки => Министерство образования и науки
соцсети => социальные сети
ОАО *«Ростелеком»* **=>** ***компания*** *«Ростелеком»*
НИУ *«Высшая Школа Экономики»* **=>** ***университет*** *«Высшая Школа Экономики»*

c) a noun, formed through the substantivisation of the past participle, is to be replaced by the construction «тот, кто + verb»

желающим бросить курить **=> тем, кто хочет** бросить курить

Similar rules have been established for all parts of speech.

It is not possible to achieve the desired results immediately when using automatic adaptation,. Given the fact that the automation process of linguistic rules and their further verification may take quite a long time, we decided to use an alternative way to find difficult constructions and simplify them. The methodological idea is to mark morphological units, which complicates the syntactical structure of the sentence. At this stage, the program looks through the text sequentially in different modes, and at each stage complex units are marked with a certain color. Then, the changes are memorized and the program returns to the homogeneous black text. Visual highlighting facilitates manual adaption, which is still the most useful way of simplifying the text:

1. The program highlights a chain of nouns in the genitive form

По словам представителя министерства, по сообщению агентства новостей, по результатам расследования фактов нарушений порядка проведения единого государственного экзамена …
о необходимости принятия мер административного характера.

2. The program marks structures consisting of a verb in the indicative mood and the infinitive if there is no punctuation. They can often be replaced by a single verb.

Позволят предотвратить, позволяет просчитать, *может привлечь*

3. Participles can be replaced by clauses later, which are easier to understand.

читавший в детстве – который читал в детстве, содержащих информацию – которые содержат информацию

4. A variety of composite conjunctions can be successfully replaced by more commonly used simple ones; the list of conjunctions is finite

не только – но и; как – так и; до тех пор, пока; несмотря на то, что and so on

Alongside the extant work on text adaptation, we worked on simplifying the syntactic structures of Russian. We began with analyzing Russian language grammar in accordance with normative reference books and textbooks, as well as courses of lectures on morphology and Russian syntax. Then we compared the material with the standards for learning Russian as a foreign language at an elementary level and made a list of what ought to be present in the

grammatical minimum. Amongst that list were syntactical and semantic complex structures, as well as more complex sentences which are too difficult to understand at an elementary level.
We divided complex structures into several categories according to Russian's syntactic structure:

- communicative: modality, emotional coloring (interjections, parenthesis, addressing)
- structural: any participles, conjunction constructions
- structural-semantic

It is important that a complicating component can be expressed with any language unit; a separate word form (often accented using particles or conjunctions); compound range; phrases; (grammatical construction of the main word and dependent words) or sentences.

Complex sentences were classified according to grammatical and formal indicators, such as the presence of several verbs with one grammatical form in the sentence. After classification, we worked on filling the classification slots and directly making rules that would exclude complex constructions that are not included in the basic language minimum.

One of the most important tasks was to describe the collection of rules within a syntactic minimum of Russian as a foreign language (RLF) for the first certification level, as well as continuously replenishing the collection of the prohibition rules. These rules describe structures that should not be present in the adapted sentences.

An example of a prohibition rule is as follows:
There should not be structures with the formula:

-the first sentence contains more than 8 wordforms, and then a compound or a complex sentence
-more than eight word forms in the sentence, then a conjuction [и а, но, или] and the second part of the compound or a complex sentence of more than 8 wordforms.

Sentences containing subjective modality words [вообразите, вообразите себе, вообще, вообще говоря, вообще-то].

The task of the linguist was to create a collection of rules which could be programmed and used as a key:
a) to extract the simplest structures from a text material of different types of complexity by replacing things that are difficult to understand and that are not included in the grammar minimum
b) to simplify complex structures that are not included in the grammar minimum

After studying the grammar guides and grammatical minimum for learning Russian, in accordance with the requirements for RLF training, we described several kinds of simple sentences (one to five word sentences). For each type of two and three-word sentences, we described the possible combinations of the sentence members' morphological compatibility.

1. ADJF masc/femn/neut, sing/plur, nomn + NOUN anim/inan, masc/femn/neut, sing/plur, nomn **Короткий срок.**

2. NOUN anim/inan, masc/femn/neut, sing/plur, nomn + NOUN anim/inan, masc/femn/neut, sing/plur, nomn **Твои вещи.**

3. NUMR masc/femn/neut, sing/plur + NOUN anim/inan, masc/femn/neut, sing/plur, nomn **Десятый билет.**

4. PRTF masc/femn/neut, sing/plur, nomn + NOUN anim/inan, masc/femn/neut, sing/plur, nomn **Стареющая кокетка.**

5. Supr masc/femn/neut, sing/plur, nomn + NOUN anim/inan, masc/femn/neut, sing/plur, nomn **Прекраснейшая девушка.**

We attempted to define a way of building an automatic algorithm to develop linguistic rules for simplifying sentences (to use materials in the electronic manual). We chose whether this was a way of excluding difficult structures which were not included in the grammar minimum, or whether it would be an algorithm based on grammatical similarity. If we had the model structures which were easy to understand then the program locates similar ones. It was decided to integrate two possible ways of achieving the best results.

5. The Lexical Adaptation Algorithm

In many cases, only using lexical adaptation methods can significantly improve the readability of the text. Moreover, these methods are relatively easy to automate, in comparison to structural adaptation. Let us consider a 'difficult' word to be replaced, *w*. We have formulated the task as compiling a list of words to replace *w* in the text, each of which has its own weight $\mathbf{R} = \{r_1, r_2 \ldots r_{S_w}\}$. The weight of *w*, which is also added to the list, is r_0.

$$\mathbf{R} = \{r_0, r_1, r_2 \ldots r_{S_w}\} \quad (1)$$

The number of word substitutes S_w depends on *w* itself. Weight r_i should reflect both ease and semantic proximity, so to calculate it we factor in the following:

1. Whether the word is included in the B1 (our target level) lexical minimum (Andriushina, 2011) – r_{i1}.

2. General word frequency in Russian – r_{i2}. This is determined using data from the Russian National Corpus[6] which consists of over 300 million words

3. Whether the word is present in the dictionary of synonyms. The dictionary contains over 300,000 words and expressions and relies on the ASIS word database (Trishin, 2010) and the AOT morphological dictionaries (Sokirco, 2004) – r_{i3}.

4. Whether the word is a hypernym or a hyponym of *w* – r_{i4}. This was taken from the YARN project (Ustalov, 2014).

5. Contextual proximity (being used in similar contexts) of the substitute under consideration and *w* – r_{i5}.

We attempt to determine the weights of the word according to each of these factors (Karpov, 2014), and then calculate the overall weight using the following formula:

$$r_i = r_{i1} \times r_{i2} \times (r_{i3} + r_{i4}) \times r_{i5} \quad (2)$$

The weights of the lexical minimum dictionary r_{i1} and dictionaries r_{i3} and r_{i4} are binary and are equal to 1 if the word is on the list and 0 if not. The contextual proximity of word *w* with itself – r_{05} is calculated as a maximum proximity value with other words.

$$r_{05} = \max_{i=1\ldots S_w} \{r_{i5}\}$$

The overall weights r_i are ranked in descending order, so that the first word on the list is the one with the greatest weight.

$$Sub = \arg\max\left(\mathbf{R} = \{r_0, r_1, r_2 \ldots r_{S_w}\}\right)$$

We consider *Sub* to be the best substitute candidate. Words with zero weight are discarded. Therefore, the suggested substitute list only contains words that are included in the lexical minimum and are in the synonyms and/or hypernyms/hyponyms dictionary. The word with the greatest weight replaces *w*.

Lexical substitution often necessitates morphological alterations to the dependent words in synthetic languages like Russian. For example, if we were to replace the rarer word *автомобиль* 'automobile' with the more frequent *машина* 'car', we would have to take into account the fact that the former is masculine, while the latter is feminine. If the original word

[6] http://www.ruscorpora.ru/en/index.html

were to be used with an attribute, e.g. *дорогой* 'expensive', we would have to change the form of the attribute, too, from masculine to feminine (→ *дорогая* 'expensive'). Word stemming and morphological processing is performed using the open application Pymorphy[7] which is based on OpenCorpora (Bocharov et al., 2011).

Using a large collection of texts of the same genre would allow us to investigate the contextual proximity of words and word groups. These data can be used in several ways. One application is measuring r_{i5}, i.e. ranking words from the dictionary of synonyms according to their contextual relevance.

The word we analyze is *w*. We choose the size of the context we are interested in and designate it as *m*. The size of the n-gram for analysis is then 2m+1=n.

I	drove	to	work	this	morning.
	-2	-1	0	1	2

Picture 2. An example of a context for *work*, m=2, n=5.

Contextual proximity can be determined by comparing the context vectors of different words. We assume that it will be a measure of their semantic similarity. The context model of an n-grams list can be represented as a hypercube in an n-dimensional space. Each axis in this model corresponds to a word in the Russian dictionary. A value on the axis is a word count in the n-gram and the fraction of a word count is a frequency.

Words context is determined using the collection *D* with 78,000 articles, most of which are news stories from the international news website Epochtimes[8]. We removed stop words and counted context frequencies near the word *w* within the n-gram. We then normalized the frequencies and removed very low values, while keeping the significant ones. As a result, each cell stores the frequency of n-gram appearance at the intersection of the corresponding values. At the preliminary stage, we built an inverted index of words in our text collection to reduce the computational complexity.

To compare the contexts of two words we ought to single out the hyperplanes of the chosen units. This is done by selecting elements w_1 and w_2 on the corresponding axes and making sections. We therefore get a subset of use frequencies of other words in the context of our word. This is a frequency vector from the context of a given width – *n*. These are to be compared and then ranked.

$$\mathbf{x}_l = \frac{NC}{N1};\ \mathbf{x}_p = \frac{NC}{N2} \tag{3}$$

where N1 is the context of the first word, N2 is the context of the second word, NC is the frequencies vector of words which are used in the contexts of both w_1 and w_2. We call this the overlapping of the contexts of both words.

There are numerous ways of calculating the distance between the resulting vectors.

We compared the overlapping of the contexts by calculating the following:

1. Euclidian distance

$$L_{Eu} = \sqrt{(\mathbf{x}_l - \mathbf{x}_p)(\mathbf{x}_l - \mathbf{x}_p)^T} \tag{4}$$

2. A cosine similarity (Tuomo Korenius et al., 2007)

$$L_{\cos} = 1 - \frac{\mathbf{x}_l \times \mathbf{x}_p^T}{\sqrt{(\mathbf{x}_l \times \mathbf{x}_l^T)(\mathbf{x}_p \times \mathbf{x}_p^T)}} \tag{5}$$

Another method for context analysis uses the Distributional Semantic Model (DSM) (Turney, 2006; Baroni and Lenci, 2010). The Latent Dirichlet Allocation (Blei et al., 2003) is a generative model that uses latent groups to explain results of observations, particularly data

[7] https://github.com/kmike/pymorphy2
[8] http://www.epochtimes.ru

similarity. For instance, if observations are words in documents, it could be said that each document is a combination of a small number of topics and that each word in the document is linked to one of the topics. The Latent Dirichlet Allocation (LDA) is one of the topic-modeling methods and was first introduced by its authors as a graphical model for topic detection.

Based on word probability distribution for topics $P(w_i / z_k);\ i \in \overline{1,|W|}, k \in \overline{1,|K|}$, we build a probability vector that corresponds to each topic. The length of this vector is equal to the number of topics *K*.

$$P(z_k / w_I);\ k \in \overline{1,|K|} \tag{6}$$

We rank the cloud of 'similar' words from the dictionary of synonyms according to the contextual distance between these words and the original one, to create a weighted cloud. The contextual distance is calculated using four different methods. We use the Kullback-Leibler divergence, as well as the Euclidian (4) and cosine (5) distances:

$$KL(p(z_k / w_{I_1}), p(z_k / w_{I_2})) = \\ = \sum_{k=1}^{K} p(z_k / w_{I_1}) \log\left(\frac{p(z_k / w_{I_1})}{p(z_k / w_{I_2})}\right) \tag{7}$$

and Jensen-Shannon divergence:

$$JS(p(z_k / w_{I_1}), p(z_k / w_{I_2})) = \\ = \frac{1}{2}\left(KL(p(z_k / w_{I_1}), \overline{p}) + KL(p(z_k / w_{I_2}), \overline{p})\right) \\ \overline{p} = \frac{1}{2}(p(z_k / w_{I_1}), p(z_k / w_{I_2})) \tag{8}$$

Since in this case we are comparing two functions of probability distribution, the divergences (7) and (8) can be easily interpreted. The calculation results for the synonyms set for the word правительство (government) is presented in Table 3.

Table 3. Context distances between the word правительство (government) and its synonyms.

Synonym	Euclid x0.01	Cosinus	KL x0.01	JS x0.01
власть vlast' 'authority'	1. 5493	0. 41598	1. 73546	0. 8771
администрация administraciya 'administration'	1. 2175	0. 67216	1. 96434	1. 1365
центр center 'center'	1. 7214	0.82965	2. 52262	2. 1914
аппарат apparat 'apparat'	1. 9592	0.98475	1. 27487	1. 7923

As can be seen from Table 3, the word *власть (authority)* has the lowest distance values for the Cosinus and Jensen-Shannon metrics.

A third similarity measure method is based on manually-crafted lexico-syntactic patterns. A paper by (Panchenko et al., 2012) shows that this measure gives results which are comparable to the baselines without the need for any fine-grained semantic resources such as WordNet. Evaluation with human judgments achieves a correlation of up to 0.739.

The lexico-syntactic similarity measure for Russian words was counted using three collections provided by Panchenko (2012). It consists of news stories from the internet and articles from Russian Wikipedia. Words with the highest similarity measure are shown in Table 4.

Table 4. Words with highest lexico-syntactic similarity measure with the word «правительство»

Wiki	Wiki+Web	Web
министерство 0.0023750842 премьер-министр 0.0020587394 национализация 0.0016480833 визирь 0.0016240252 виконт 0.0015276774 министр 0.0015159135 парламент 0.0014622179 автономия 0.0013840345 канцелярия 0.0013651757	министерство 0.0032323967 парламент 0.0028992076 корпорация 0.0025010496 дума 0.0018373777 вице-премьер 0.0014422634 фмс 0.0013199818 администрация 0.0012130674 спекулянт 0.0011987228	корпорация 0.0121555394 министерство 0.0093719878 вице-премьер 0.0088692562 парламент 0.0081208092 фмс 0.0078992850 спекулянт 0.0069952559 салех 0.0042232399 медиакампания 0.0041847118 сексменьшинство 0.0040592306 распоряжение 0.0040407077 минфин 0.0039092948

For this purpose we used similarity calculated by the joint collection (Wiki+Web) as the more common model.

6. Adaptation case study

The adaptation process for one text fragment includes the following steps during preprocessing:

- tokenize the text
- lemmatize the tokens
- index the words using the lemmas dictionary

In the YARN project, the first word, **исследование** (research) has 10 synsets which includes 30 synonyms. For instance: **изучение** 'study', **пытка** 'torture', **суд** 'judgement', **эксперимент** 'experiment'. Only 5 of these 30 are also a part of the B1 level's active vocabulary: **анализ** 'analysis', **труд** 'work', **эксперимент** 'experiment', **опыт** 'experience', **книга** 'book'. The semantic similarity of the target word with substitutions, as well as other weights, is shown in Table 5.

Table 5. Weights of substitutions for the word *исследование* (research)

Synonym	r_{i1}	r_{i2}	r_{i3}	r_{i4}	r_{i5}	r_i
анализ 'analysis'	1	19061	1	0	0.0110236631	135.749
труд 'work'	1	55886	1	0	0.0000854651	5.465
эксперимент 'experiment'	1	7862	1	0	0.0006164419	3.803
опыт 'experience'	1	40981	1	0	0.0001583697	3.766
книга 'book'	1	79314	1	0	0.0000567084	3.585

The word **анализ** (analysis) has the highest similarity measure. Therefore, the system selects it as a substitute for **исследование** (research). Unfortunately, this approach cannot be extended to other parts of speech such as verbs, because at present we do not have a proven contextual proximity ser for Russian verbs.

The result of applying the developed algorithm is shown below:

Специалисты *"ЛК" провели* ***анализ*** *трех российских социальных сетей "Вконтакте", "Одноклассники" и "Мой мир". В ходе недельного* ***анализа специалисты*** *выяснили, что ребенку в возрасте 13 лет в соцсети "Вконтакте" по* ***вопросу*** *порно выдается список из закрытых групп, но при введении в поисковый* ***вопрос*** *синонимов этого слова ребенок получает перечень открытых групп содержащих информацию по запрашиваемой тематике и может без* ***вопросов*** *ее просмотреть. Кроме того (страна) ребенок не застрахован от общения с педофилами в соцсети, может подвергаться моральному насилию и запугиванию, а также столкнуться с мошенниками. В социальной сети*

*"Одноклассники" по **вопросу** порно и его синонимов ребенку становится доступен список закрытых группах в которые нельзя получить доступ. В этой социальной сети **ребёнок** может рассчитывать только на "легкую эротику". Общение с педофилами в "Одноклассниках" сведено к минимуму а каких либо видов мошенничества **специалисты** в сети не обнаружили.*

Evidently, the developed approach can find substitutes for some words in the text without losing the main meaning. We explore this approach for lexical adaptation using a set of 10 texts which were mentioned Section 4. The analysis indicates satisfactory results. A more precise evaluation requires larger datasets and blind assessors meaning, which we are planning to undertake and expand upon in future research.

7. Conclusion

The creation of an automatic system for retrieving texts which are appropriate for educational purposes is a practically-oriented investigative activity. It opens up new horizons not only in education, but also in the commercial sector, such as rewriting.

The classical models which had been developed for the prediction of English language readability were adapted to Russian. These models, as well as others, were developed specially for Russian and were tested on our data. The accuracy of four levels of classification with the Random Forest methods reached 98-99%. Given this, we could say that results met our needs. We managed to develop a precise classification system for the readability of news texts in Russian. The best approach for single sentence readability prediction was the Random Forest approach, which gave a classification accuracy of 0.94.

An algorithm for the lexical adaptation of news articles which can be used as materials for learning and teaching Russian as a foreign language is described in this paper in detail. The algorithm relies on the following substitute-ranking factors:

1. Whether the word is included in the B1 (our target level) lexical minimum (Andriushina, 2011);
2. General word frequency in Russian and in texts of the selected genre
3. Whether the word is present in the dictionary of synonyms
4. Whether the word is a hypernym or a hyponym of *w*
5. Contextual proximity of the substitute under consideration to *w*

We considered three methods of calculating contextual proximity. The first relies on the vector of normalized frequencies of word use in the closest context. The second is based on the LDA model and on the vector of topic-based word frequencies distribution. We have found that in both cases contextual proximity yields useful results for ranking synonyms.

A limitation of the first method that is based on calculating term frequencies in n-grams is a high data dimension. The vector length is equal to the size of the dictionary, which is around 20,000 words in our case. The second method, based on the LDA model, solves the problem of high dimension and allows us to efficiently calculate and interpret contextual proximity.

With the LDA model, clouds of similar words and word groups for the given word can be ranked. The LDA model also allows us to generate document descriptions and find clouds of similar documents.

Acknowledgments

This study used research findings from the «Adaptation of texts from the Russian National Corpus for the electronic textbook «Russian language as a foreign one» conducted at The National Research University Higher School of Economics' Academic Fund Program in 2013, grant No 13-05-0031.

References

Andriushina, N., 2011. Russian as a Foreign Language Lexical Minimum. First Certificate Level. General Proficiency., Zlatoust. ed. Saint-Petersburg, Russia.

Baroni, M., Lenci, A., 2010. Distributional memory: A general framework for corpus-based semantics. Comput. Linguist. 36, 673–721.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res. 993–1022.

Bocharov, V., Stepanova, M., Ostapuk, N., Bichineva, S., Granovsky, D., 2011. Quality assurance tools in the OpenCorpora project, in: Computational Linguistics and Intelligent Technology: Proceeding of the International Conference «Dialog–2011». pp. 10–17.

Chandrasekar, R., Doran, C., Srinivas, B., 1996. Motivations and methods for text simplification, in: Proceedings of the 16th Conference on Computational Linguistics-Volume 2. Association for Computational Linguistics, pp. 1041–1044.

Franccois, T.L., 2009. Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. Association for Computational Linguistics, pp. 19–27.

Karpov, N., 2014. Corpus-Based Text Retrieval and Adaptation for Learning System. Int. J. Adv. Comput. Sci. Its Appl. 4, 38–43.

Karpov, N., Vitugin, F., Baranova, J., 2014. Single-sentence Readability Prediction in Russian, in: 3rd International Conference on Analysis of Images, Social Networks and Texts, Communications in Computer and Information Science.

Mackovskiy, M.S., 1973. The problem of understanding of printed texts by readers (sociological analysis). Moscow, Russia.

Nevdah, M., 2008. Development of a method of automated evaluation of the complexity of educational texts for higher school.

Nivre, J., Boguslavsky, I.M., Iomdin, L.L., 2008. Parsing the SynTagRus treebank of Russian, in: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, pp. 641–648.

Oborneva, I., 2006. Automatic assessment of the complexity of educational texts on the basis of statistical parameters. Moscow, Russia.

Panchenko, A., Morozova, O., Naets, H., 2012. A semantic similarity measure based on lexico-syntactic patterns, in: Proceedings of KONVENS. pp. 174–178.

Sokirco, A., 2004. Morphological Modules on the Site www.aot.ru, in: Computational Linguistics and Intelligent Technology: Proceeding of the International Conference «Dialog–2004».

Trishin, V., 2010. Electronic dictionary and handbook of Russian synonyms in the ASIS system, in: Vladimir Dal at Happy Home on the Presny Str. Academia, Moscow, Russia, pp. 158–165.

Tuomo Korenius, Jorma Laurikkala, Martti Juhola, 2007. On principal component analysis, cosine and Euclidean measures in information retrieval. Inf. Sci. 177, 4893–4905.

Turney, P.D., 2006. Similarity of semantic relations. Comput. Linguist. 32, 379–416.

Ustalov, D., 2014. Enhancing Russian Wordnets Using the Force of the Crowd, in: Analysis of Images, Social Networks and Texts. Springer, pp. 257–264.

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., North, B., 2009. Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge University Press.

Contact details and disclaimer:

Nikolay V. Karpov
National Research University Higher School of Economics (Nizhny Novgorod, Russia). «Department of Applied Mathematics and Informatics». Dotsent;
E-mail: nkarpov@hse.ru, Tel. +7 (950) 607-21-52

Vera G. Sibirtseva
National Research University Higher School of Economics (Nizhny Novgorod, Russia). «Department of Applied Linguistics and Intercultural Communication». Dotsent;
E-mail: vsibirtseva@hse.ru, Tel. +7 (952) 767-25-20