Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories

Robert Reynolds
UiT: The Arctic University of Norway
Postboks 6050 Langnes
9037 Tromsø, Norway
robert.reynolds@uit.no

Abstract

I investigate Russian second language readability assessment using a machine-learning approach with a range of lexical, morphological, syntactic, and discourse features. Testing the model with a new collection of Russian L2 readability corpora achieves an F-score of 0.671 and adjacent accuracy 0.919 on a 6-level classification task. Information gain and feature subset evaluation shows that morphological features are collectively the most informative. Learning curves for binary classifiers reveal that fewer training data are needed to distinguish between beginning reading levels than are needed to distinguish between intermediate reading levels.

1 Introduction

Reading is one of the core skills in both first and second language learning, and it is arguably the most important means of accessing information in the modern world. Modern second language pedagogy typically includes reading as a major component of foreign language instruction. There has been debate regarding the use of authentic materials versus contrived materials, where *authentic* materials are defined as "A stretch of real language, produced by a real speaker or writer for a real audience and designed to convey a real message of some sort" (Morrow, 1977, p. 13).[1] Many empirical studies have demonstrated advantages to using authentic materials, including increased linguistic, pragmatic, and discourse competence (Gilmore, 2007, citations in §3). However, Gilmore (2007) notes that "Finding appropriate authentic texts and designing tasks for them can, in itself, be an extremely time-consuming process." An appropriate text should arguably be interesting, linguistically relevant, authentic, recent, and at the appropriate reading level.

Tools to automatically identify a given text's complexity would help remove one of the most time-consuming steps of text selection, allowing teachers to focus on pedagogical aspects of text selection. Furthermore, these tools would also make it possible for learners to find appropriate texts for themselves.

A thorough conceptual and historical overview of readability research can be found in Vajjala (2015, §2.2). The last decade has seen a rise in research on readability classification, primarily focused on English, but also including French, German, Italian, Portuguese, and Swedish (Roll et al., 2007; Vor der Brück et al., 2008; Aluisio et al., 2010; Francois and Watrin, 2011; Dell'Orletta et al., 2011; Hancke et al., 2012; Pilán et al., 2015). Broadly speaking, these languages have limited morphology in comparison with Russian, which has relatively rich morphology among major world languages. It is therefore not surprising that morphology has received little attention in studies of automatic readability classification. One important exception is Hancke et al. (2012) which examines lexical, syntactic and morphological features with a two-level corpus of German magazine articles. In their study, morphological features are collectively the most predictive category of features. Furthermore, when combining feature categories in groups of two or three, the

[1] The definition of authenticity is itself a matter of disagreement (Gilmore, 2007, §2), but Morrow's definition is both well-accepted and objective.

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 289–300,
San Diego, California, June 16, 2016.

highest performing combinations included the morphology category. If morphological features figure so prominently in German readability classification, then there is good reason to expect that they will be similarly informative for Russian second-language readability classification.

This article explores to what extent textual features based on morphological analysis can lead to successful readability classification of Russian texts for language learning. In Section 2, I give an overview of previous research on readability, including some work on Russian. The corpora collected for use in this study are described in Section 3. The features extracted for machine learning are outlined in Section 4. Results are discussed in Sections 5 and 6, and conclusions and outlook for future research are presented in Section 7.

2 Background

The history of empirical readability assessment began as early as 1880 (DuBay, 2006), with methods as simple as counting sentence length by hand. Today, research on readability is dominated by machine-learning approaches that automatically extract complex features based on surface wordforms, part-of-speech analysis, syntactic parses, and models of lexical difficulty. In this section, I give an abbreviated history of the various approaches to readability assessment, including the kinds of textual features that have received attention. Although some proprietary solutions are relevant here, I focus primarily on work that has resulted in publically available knowledge and resources.

2.1 History of evaluating text complexity

The earliest approaches to readability analysis consisted of developing readability formulas, which combined a small number of easily countable features, such as average sentence length, and average word length (Kincaid et al., 1975; Coleman and Liau, 1975). Although formulas for computing readability have been criticized for being overly simplistic, they were quickly adopted and remain in widespread use today.[2] An early extension of these simple 'counting' formulas was to additionally rely on lists of words deemed "easy", based on either their frequency or polling of young learners (Dale and Chall, 1948; Chall and Dale, 1995; Stenner, 1996). A higher proportion of words belonging to these lists resulted in lower readability measures, and vice versa.

With the recent growth of natural language processing techniques, it has become possible to extract information about the lexical and/or syntactic structure of a text, and automatically train readability models using machine-learning techniques. Some of the earliest attempts at this built unigram language models based on American textbooks, and estimated a text's reading level by testing how well it was described by each unigram model (Si and Callan, 2001; Collins-Thompson and Callan, 2004). This approach was extended in the REAP project[3] to include a number of grammatical features as well (Heilman et al., 2007; Heilman et al., 2008a; Heilman et al., 2008b).

Over time, readability researchers have increasingly taken inspiration from various subfields of linguistics to identify features for modeling readability, including syntax (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009), discourse (Feng, 2010; Feng et al., 2010), textual coherence (Graesser et al., 2004; Crossley et al., 2007a; Crossley et al., 2007b; Crossley et al., 2008), and second language acquisition (Vajjala and Meurers, 2012). The present study expands this enterprise by examining second language readability for Russian.

2.2 Automatic readability assessment of Russian texts

The history of readability assessment of Russian texts takes a very similar trajectory to the work related above. Early work was based on developing formulas based on simple countable features (Mikk, 1974; Oborneva, 2005; Oborneva, 2006a; Oborneva, 2006b; Mizernov and Graščenko, 2015).

Some researchers have tried to be more objective about defining readability, by obtaining data from expert raters, or from other experimental means, and then performing statistical analysis—such as linear regression, or correlation—to identify impor-

[2] The Flesch Reading Ease test and the Flesch-Kincaid Grade Level test are implemented in the proofing tools of many major word processors.

[3] http://reap.cs.cmu.edu

tant factors of text complexity (Sharoff et al., 2008; Petrova and Okladnikova, 2009; Okladnikova, 2010; Špakovskij, 2003; Špakovskij, 2008; Ivanov, 2013; Kotlyarov, 2015), such as lexical properties, morphological categories, typographic layout, and syntactic complexity.

To my knowledge, only one study has previously examined readability in the context of Russian second-language pedagogical texts. Karpov et al. (2014) performed a series of experiments using several different kinds of machine-learning models to automatically classify Russian text complexity, as well as single-sentence complexity. They collected a small corpus of texts (described in Section 3 below), with texts at 4 of the CEFR levels:[4] A1, A2, B1, and C2. They extracted 25 features from these texts, including document length, sentence length, word length, lexicon difficulty, and presence of each part of speech. No morphological features were included, despite the fact that morphology is the most challenging feature of Russian grammar for most language learners. Using Classification Tree, SVM, and Logistic Regression models for binary classification (A1-C2, A2-C2, and B1-C2), they report achieving accuracy close to 100%. It should be noted that no results were reported with more customary stepwise binary combinations, such as A1-A2, A2-B1, and B1-C2, which are more difficult—and more useful—distinctions. In a four-way classication task, they state that their results were lower, but they only provide precision, recall, and accuracy metrics for the B1 readability level during four-way classification, which were as high as 99%. Irregularities in reporting make it difficult to draw firm conclusions from their work, especially because their corpora covered only four out of six CEFR levels with no more than 60 data points per level.

3 Corpora

The corpora[5] in this study all use the same scale for rating L2 readability, the Common European Framework of Reference for Languages (CEFR). The six common reference levels of CEFR can be divided into three levels—Basic user (A), Independent user (B), and Proficient user (C)—each of which is subdivided into two levels. This yields the following six levels in ascending order: A1, A2, B1, B2, C1, and C2.[6] For all corpora, reading levels were assigned by the original author or publisher, so there is no guarantee that the reading levels between corpora align well.

Two subcorpora were used by Karpov et al. (2014). The CIE corpus includes texts created by teachers for learners of Russian. These texts are taken from a collection of materials kept in an open repository at `http://texts.cie.ru`. The second subcorpus used by Karpov et al. (2014) consists of 50 original news articles for native readers, rated at level C2.

The LingQ corpus (LQ) is a corpus of texts from `http://www.lingq.com`, a commercial language-learning website that includes lessons uploaded by member enthusiasts, with 3481 texts. Reading levels were determined by the member who uploaded each lesson.

The Red Kalinka (RK) corpus is a collection of 99 texts taken from 13 books in the "Russian books with audio" series available at `http://www.redkalinka.com`. These books include stories, dialogues, texts about Russian culture, and business dialogues.

The TORFL corpus comes from the Test of Russian as a Foreign Language, a set of standardized tests administered by the Russian Ministry of Education and Science. It is a collection of 168 texts that I extracted from official practice tests for the TORFL.

The Zlatoust corpus (Zlat) comes from a series of readers for language learners at the lower CEFR levels, with 746 documents.

The Combined corpus is a combination of the corpora described above. The distribution of documents per level is given in Table 1. Note that some corpora do not have texts at every reading level.

Table 2 shows the median document length (in words) per level in each of the corpora. The overall median document size is 268 words. Within each corpus, median document length tends to in-

[4] CEFR levels are introduced in Section 3.

[5] Some of the corpora used in this study are proprietary, so they cannot be published online. However, they can be shared privately for research purposes. With the exception of the two corpora from Karpov et al. (2014), all of the corpora were created and used for the first time in this study.

[6] There is no consensus on how the CEFR levels align with other language evaluation scales, such as the ACTFL and ILR used in the United States.

	All	A1	A2	B1	B2	C1	C2
CIE	145	28	57	60	–	–	–
news	50	–	–	–	–	–	50
LQ	3481	323	653	716	832	609	348
RK	99	40	18	17	18	6	–
TORFL	168	31	36	36	26	28	11
Zlat.	746	–	66	553	127	–	–
Comb.	4689	422	830	1382	1003	643	409

Table 1: Distribution of documents per level for each corpus

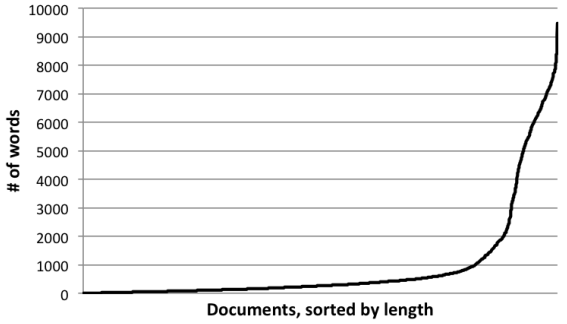
crease with each level, with some exceptions. Tests were conducted with a modified corpus in which longer documents were truncated to approximately 300 words; classifier performance was slightly lower with this modified corpus.

	All	A1	A2	B1	B2	C1	C2
CIE	314	116	340	354	–	–	–
news	174	–	–	–	–	–	174
LQ	246	65	47	225	522	3247	436
RK	286	68	296	418	278	292	–
TORFL	158	55	160	196	238	146	284
Zlat.	344	–	122	345	414	–	–
Comb.	268	67	68	275	474	2621	313

Table 2: Median words per document for each level of each corpus

The overall distribution of document length is shown in Figure 1, where the x-axis is all documents ranked by document length and the y-axis is document length. The shortest document contains 7 words, and the longest document contains over 9000 words.

Figure 1: Distribution of document length in words



4 Features

In the following sections, I give an overview of the features used in this study, both the rationale for their inclusion, as well as details regarding their operationalization and implementation. I combine features used in previous research with some novel features based on morphological analysis. I divide features into the following categories: lexical, morphological, syntactic, and semantic.

4.1 Lexical features (LEX)

The lexical features (LEX) are divided into three subcategories: lexical variability (LEXV), lexical complexity (LEXC), and lexical familiarity (LEXF).

LEXV The lexical variability category contains features that are intended to measure the variety of lexemes found in a document. One of the most basic measures of lexical variability is the type-token ratio, which is the number of unique wordforms divided by the number of tokens in a text. Because the type-token ratio is dependent on document length, I included a few more robust metrics that have been proposed: Root TTR ($T/\sqrt{N}$), Corrected TTR ($T/\sqrt{2N}$), Bilogarithmic TTR ($\log T/\log N$), and the Uber Index ($\log^2 T/\log(N/T)$). For all of these metrics, a higher score signifies higher concentrations of unique tokens, which indicates more difficult readability levels.

LEXC Lexical complexity includes multiple concepts. One is the degree to which individual words can be parsed into component morphemes. This is a reflection of the derivational or agglutinative structure of words. Another measure of lexical complexity is word length, which reflects the difficulty of chunking and storing words in short-term memory. Depending on the particulars of a given language or the development level of a given learner, lexical complexity can either inhibit or enhance comprehension. For example, the word *neftepererabatyvajuščij (zavod)* 'oil-refining (factory)' is overwhelming for a beginning learner, but an advanced learner who has never seen this word can easily deduce its meaning by recognizing its component morphemes: *nefte-pere-rabat-yvaj-uščij* 'oil-re-work-IPFV-ing'.

Word length features were computed on the basis of characters, syllables, and morphemes. For each of these three, both an average and a maximum were computed. In addition, all six of these features were computed for both all words, and for content

words only.[7] The features for word length in morphemes were computed on the basis of Tixonov's Morpho-orthographic dictionary (Tixonov, 2002), which contains parses for about 100 000 words. All words that are not found in the dictionary were ignored. In addition to average and maximum word lengths, I also followed Karpov et al. (2014) in calculating word length bands, such as the proportion of words with five or more characters. These bands are calculated for 5–13 characters (9 features) and 3–6 syllables (4 features). All 13 of these features were calculated both for all words and for content words only.

LEXF Lexical familiarity features were computed to attempt to capture the degree to which the words of a text are familiar to readers of various levels. These features model the development of learners' vocabulary from level to level. Unlike the features for lexical variability and lexical complexity, which are primarily based on surface structure, the features for lexical familiarity rely on a predefined frequency lists or lexicons.

The first set of lexical familiarity features are derived from the official "Lexical Minimum" lists for the TORFL examinations. The lexical minimum lists are compiled for the four lowest levels (A1, A2, B1, and B2), where each list contains the words that should be mastered for the tests at each level. These lists can be seen as *prescriptive* vocabulary for language learners. Following Karpov et al. (2014), I computed features for the proportion of words above a given reading level.

The second set of lexical familiarity features are taken from the Kelly Project (Kilgarriff et al., 2014), which is a "corpus-based vocabulary list" for language learners. These lists are based primarily on word frequency, with manual adjustments made by professional teachers. Just like the features based on the Lexical Minimum, I computed the proportion of words over each of the six CEFR levels.

The third set of lexical familiarity features are based on raw frequency and frequency rank for both lemma frequency and token frequency.[8] For each of the four kinds of frequency data, I computed average, median, minimum, and standard deviation.

4.2 Morphological features (MORPH)

Morphological features are primarily based on morphosyntactic values, as output by an automatic morphological analyzer. The first three sets of features reflect simple counts of whether a morphosyntactic tag is present or what proportion of tokens receive each morphosyntactic tag. The first set of features expresses whether a given morphosyntactic tag is present in the document. A second set of features, expresses the ratio of tokens with each morphosyntactic tag, normalized by token count. A third set of features, the value-feature ratio (VFR), was calculated as the number of tokens that express a morphosyntactic value (e.g. past), normalized by the number of tokens that express the corresponding morphosyntactic feature (e.g. tense).

In the early stages of learning Russian, learners do not have a knowledge of all six cases, so I hypothesized that texts intended for the lowest reading level might be distinguished by a limited number of attested cases. Similarly, two subcases in Russian, partitive genitive and second locative, are generally rare, but are overrepresented in texts written for beginners who are being introduced to these subcases. Two features were computed to capture these intuitions: the number of cases and the number of subcases attested in the document.

Following Nikin et al. (2007; Krioni et al. (2008; Filippova (2010), I calculated a feature to measure the proportion of abstract words. This was done by using a regular expression to test lemmas for the presence of a number of abstract derivational suffixes. This feature is normalized to the number of tokens in the document.

4.2.1 Sentence length-based features (SENT)

The SENT category consists of features that include in their computation some form of sentence length, including words per sentence, syllables per sentence, letters per sentence, coordinating conjunctions per sentence, and subordinating conjunc-

[7] The following parts of speech were considered content words: adjectives, adverbs, nouns and verbs.

[8] Lemma frequency data were taken from Ljaševskaja and Šarov (2009) (available digitally at `http://dict.ruslang.ru/freq.php`), which is based on data from the Russian National Corpus. The token frequency data were taken directly from the Russian National Corpus webpage at `http://ruscorpora.ru/corpora-freq.html`.

tions per sentence. In addition, I also compute the type frequency of morphosyntactic readings per sentence. This category also includes the traditional readability formulas: Russian Flesch Reading Ease (Oborneva, 2006a), Flesch Reading Ease, Flesch-Kincaid Grade Level, and the Coleman-Liau Index.

4.3 Syntactic features (SYNT)

Syntactic features for this study were primarily based on the output of the `hunpos`[9] trigram part-of-speech tagger and `maltparser`[10] syntactic dependency parser, both trained on the SynTagRus[11] treebank. Using `maltoptimizer`,[12] I found that the best-performing algorithm was Nivre Eager, which achieved a labeled attachment score of 81.29% with cross-validation of SynTagRus.

Researchers of automatic readability classification and closely related tasks have used a number of syntactic dependency features which I also implement here (Yannakoudakis et al., 2011; Dell'Orletta et al., 2011; Vor der Brück and Hartrumpf, 2007; Vor der Brück et al., 2008). These include features based on dependency lengths (the number of tokens intervening between a dependent and its head), as well as the number of dependents belonging to particular parts of speech, in particular nouns and verbs. In addition, I also include features based on dependency tree depth (the path length from root to leaves).

4.4 Discourse/content features (DISC)

The discourse/content features (DISC) are intended to capture the broader difficulty of understanding the text as a whole, rather than the difficulty of processing the linguistic structure of particular words or sentences. One set of features are based on *definitions* (Krioni et al., 2008), which are a set of words and phrases that are used to introduce or define new terms in a text. Using regular expressions, I calculate definitions per token and definitions per sentence.

Another set of features is adapted from the work of Brown et al. (2007; 2008), who show that logical propositional density—a fundamental measurement in the study of discourse comprehension—can be accurately measured purely on the basis of part-of-speech counts.

One other feature is based on the intuition that reading dialogic texts is generally easier than reading prose. This feature is computed as the number of dialog symbols[13] per token.

4.5 Summary of features

As outlined in the preceding sections, this study makes use of 179 features. Many of the features are inspired by previous research of readability, both for Russian and for other languages. The distribution of these features across categories is shown in Table 3.

Category	Number of features
DISC	6
LEXC	42
LEXF	38
LEXV	7
MORPH	60
SENT	10
SYNT	16
Total	179

Table 3: Distribution of features across categories

5 Results

The machine-learning and evaluation for this study were performed using the `weka` data mining software (Hall et al., 2009). Based on preliminary tests, the Random Forest model was selected as the classifier algorithm for the study.[14] All results reported below are achieved using the Random Forest algorithm with default parameters. Unless otherwise specified, evaluation was performed using ten-fold cross validation.

Results are given in Table 4. *Precision* is a measure of how many of the documents predicted to be at a given readability level are actually at that level (true positives divided by true and false positives).

[9] `https://code.google.com/p/hunpos/`

[10] `http://www.maltparser.org/`

[11] `http://ruscorpora.ru/instruction-syntax.html`

[12] `http://nil.fdi.ucm.es/maltoptimizer/index.html`

[13] In Russian, -, –, — and : are used to mark turns in a dialog.

[14] Other classifiers that consistently performed well were NNge (nearest-neighbor with non-nested generalized exemplars), FT (Functional Trees), MultilayerPerceptron, and SMO (sequential minimal optimization for support vector machine).

Recall measures how many of the documents at a given readability level are predicted correctly (true positives divided by true positives and false negatives). The two metrics are calculated for each reading level and weighted averages are reported for the classifier as a whole. The *F-score* is a harmonic mean of precision and recall. *Adjacent accuracy* is the same as weighted recall, except that it considers predictions that are off by one category as correct. For example, a B2 document is counted as being correctly classified if the classifier predicts B1, B2, or C1. The baseline performance achieved by predicting the mode reading level (B1)—using `weka`'s ZeroR classifier—is precision 0.097 and recall 0.312 (F-score 0.149). The OneR classifier, which is based on only the most informative feature (corrected type-token ratio), achieves precision 0.487 and recall 0.497 (F-score 0.471). The Random Forest classifier, trained on the full Combined corpus with all 179 features, achieves precision 0.69 and recall 0.677 (F-score 0.671), with adjacent accuracy 0.919.

Classifier	Precis.	Recall	F-score
ZeroR	0.097	0.312	0.149
OneR	0.487	0.497	0.471
RandomForest	0.690	0.677	0.671

Table 4: Baseline and RandomForest results with Combined corpus

A confusion matrix is given in Table 5, which shows the predictions of the RandomForest classifier. The rows represent the actual reading level as specified in the gold standard, whereas the columns represent the reading level predicted by the classifier. Correct classifications appear along the diagonal. Table 5 shows that the majority of misclassifications are only off by one level, and indeed the adjacent accuracy is 0.919, which means that less than 10% of the documents are more than one level away from the gold standard.

5.1 Binary classifiers

Evaluation was performed with binary classifiers, in which the datasets contain only two adjacent readability levels. Since the Combined corpus has six levels, there are five binary classifier pairs: A1-A2, A2-B1, B1-B2, B2-C1, C1-C2. The results of the cross-validation evalution of these classifiers is given in Table 6. Red Kalinka and LQsupp (the second largest subcorpus of LingQ)—which were judged to be the most reliable subcorpora—were also examined individually.

	A1	A2	B1	B2	C1	C2
A1	**234**	120	48	0	0	0
A2	41	**553**	192	17	0	0
B1	16	76	**1130**	90	5	5
B2	1	57	311	**478**	83	4
C1	1	20	66	98	**394**	6
C2	0	3	40	58	9	**78**

Table 5: Confusion matrix for RandomForest, all features, Combined corpus. Rows are actual and columns are predicted.

		A1-A2	A2-B1	B1-B2	B2-C1	C1-C2
Comb.	prec.	0.821	0.857	0.817	0.833	0.894
	recall	0.821	0.857	0.811	0.831	0.897
	F-score	0.812	0.855	0.806	0.826	0.892
RK	prec.	0.967	0.943	0.832	0.837	–
	recall	0.966	0.943	0.829	0.792	–
	F-score	0.965	0.943	0.828	0.730	–
LQsupp	prec.	0.911	0.806	0.955	0.914	0.926
	recall	0.903	0.806	0.956	0.915	0.924
	F-score	0.901	0.806	0.954	0.912	0.924

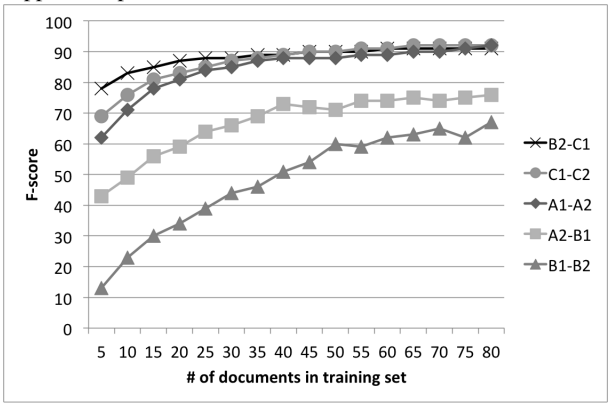
Table 6: Evalution metrics for binary classifiers: RandomForest, all features

As expected, because the binary classifiers' are more specialized, with less data noise and fewer levels to choose between, their accuracy is much higher.

One potentially interesting difference between binary classifiers at different levels is their learning curves, or in other words, the amount of training data needed to approach optimal results. I hypothesized that the binary classifiers at lower levels would need less data, because texts for beginners have limited possibilities for how they can vary without increasing complexity. Texts at higher reading levels, however, can vary in many different ways. To adapt Tolstoy's famous opening line to *Anna Karenina*, "All [simple texts] are similar to each other, but each [complex text] is [complex] in its own way." If this is true, then binary classifiers at higher reading levels should require more data to reach the upper limit of their classifying accuracy. This prediction was tested by controlling the number of documents used in the training data for each binary classifier, while tracking the F-score on cross-validation. Re-

sults of this experiment are given in Figure 2.

Figure 2: Learning curves of binary classifiers trained on LQ-supp subcorpus



The results of this experiment support the hypothesized difference between binary classifier levels, albeit with some exceptions. The A1-A2 classifier rises quickly, and begins to level off after seeing about 40 documents. The A2-B1 classifier rises more gradually, and levels off after seeing about 55 documents. The B1-B2 classifier rises even more slowly, and does not level off within the scope of this figure.

Up to this point, the data confirm my hypothesis that lower levels require less training data. However, the B2-C1 and C1-C2 classifiers buck this trend, with learning curves that outperform the simplest binary classifier with very little training data. One possible explanation for this is that the increasing complexity of CEFR levels is not linear, meaning that the leap from A1 to A2 is much smaller than the leap from C1 to C2. The increasing rate of change is explicitly formalized in the official standards for the TORFL tests. For example, the number of words that a learner should know has the following progression: 750, 1300, 2300, 10 000, 12 000 (7 000 active), 20 000 (8 000 active). This means that distinguishing B2-C1 and C1-C2 should be easier because the distance between their respective levels is an order of magnitude larger than the distance between the respective levels of A1-A2, A2-B1. Furthermore, development of grammar should be more or less complete by level B2, so that the the number of features that distinguish C1 from C2 should be smaller than in lower levels, where grammar development is a limiting factor.

6 Feature evaluation

As summarized in Section 4.5, this study makes use of 179 features, divided into 7 categories: DISC, LEXC, LEXF, LEXV, MORPH, SENT, and SYNT. Many of the features used in this study are taken from previous research of related topics, and some features are proposed for the first time here. Previous researchers of Russian readability have not included morphological features, so the results of these features are of particular interest here.

In this section, I explore the extent to which the selected corpora can support the relevance and impact of these features in Russian second language readability classification. One rough test for the value of each category of features is to run cross-validation with models trained on only one category of features. In Table 7, I report the results of this experiment using the Combined corpus.

Category	# features	precision	recall	F-score
DISC	6	0.482	0.482	0.477
LEXC	42	0.528	0.532	0.514
LEXF	38	0.581	0.573	0.567
LEXV	7	0.551	0.552	0.546
MORPH	60	0.642	0.627	0.618
SENT	10	0.478	0.479	0.474
SYNT	16	0.518	0.533	0.514
LEXC+LEXF+LEXV	87	0.652	0.645	0.639

Table 7: Precision, recall, and F-score for six-level Random Forest models trained on the Combined corpus

The results in Table 7 show that MORPH, has the highest F-score of any single category, with an F-score just 0.053 below a model trained on all 179 features. True comparisons between categories are problematic because the number of features per category varies significantly.

In order to evaluate the usefulness of each feature as a member of a feature set, I used the correlation-based feature subset selection algorithm (CfsSubsetEval) (Hall, 1999), which selects the most predictive subset of features by minimizing redundant information, based on feature correlation.

Out of 179 features, the CfsSubsetEval algorithm selected 32 features. Many of the features selected for the optimal feature set are also among the top 30 most informative features according to information gain. However, the morphological features—which had only 7 features among the top 30 for information

gain—now include 14 features, which indicates that although these features are not as informative, the information that they contribute is unique.

A classifier trained on only these 32 features with the Combined corpus achieved precision 0.674 and recall 0.665 (F-score 0.659), which is only 0.01 worse than the model trained on all 179 features.

7 Conclusions and Outlook

This article has presented new research in automatic classification of Russian texts according to second language readability. This technology is intended to support learning activities that enhance student engagement through online authentic materials (Erbaggio et al., 2010). I collected a new corpus of Russian language-learning texts classified according to CEFR proficiency levels. The corpus comes from a broad spectrum of sources, which resulted in a richer and more robust dataset, while also complicating comparisons between subsets of the data.

Classifier performance A six-level Random Forest classifier achieves an F-score of 0.671, with adjacent accuracy of 0.919. Binary classifiers with only two adjacent reading levels achieve F-scores between 0.806 and 0.892. This is the first large-scale study of this task with Russian data, and although these results are promising, there is still room for improvement, both in corpus quality and modeling features.

In Section 5.1, I showed that binary classifiers at the lowest and highest reading levels required less training data to approach their upper limit. Beginning with the lowest levels, each successive binary classifier learned more slowly than the last until the B2-C1 level. I interpret this as evidence that simple texts are all similar, but complex texts can be complex in many different ways.

Features Among the most informative individual features used in this study are type-token ratios, as well as various measures of maximum syntactic dependency lengths and maximum tree depth. However, as a category, the morphological features are most informative. When features with overlapping information are removed using correlation-based feature selection, the resulting set includes 14 MORPH features, 8 SYNT features, 4 LEXV features, 3 LEXF features, and 2 LEXC features, and 1 DISC feature. Models trained on only one category of features also show the importance of morphology in this task, with the MORPH category achieving a higher F-score than other individual categories.

Although the feature set used in this study had fairly broad coverage, there are still a number of possible features that could likely improve classifier performance further. Other researchers have seen good results using features based on semantic ambiguity, derived from word nets. Implementing such features would be possible with the new and growing resources from the Yet Another RussNet project.[15]

Another category of features that is absent in this study is language modeling, including the possibility of calculating information-theoretic metrics, such as surprisal, based on those models.

The syntactic features used in this study could be expanded to capture more nuanced features of the dependency structure. For instance, currently implemented syntactic features completely ignore the kinds of syntactic relations between words. In addition, some theoretical work in dependency syntax, such as catenae (Osborne et al., 2012) and dependency/locality (Gibson, 2000) may serve as the basis for other potential syntactic features.

Applications One of the most promising applications of the technology discussed in this article is a grammar-aware search engine or similar information retrieval framework that can assist both teachers and students to identify texts at the appropriate reading level. Such systems have been discussed in the literature (Ott, 2009), and similar tools can be created for Russian language learning.

Acknowledgments

I am indebted to Detmar Meurers and Laura Janda for insightful feedback at various stages of this project. I am grateful to Nikolay Karpov for openly sharing his research source files. I am also thankful to the CLEAR research group at UiT and three anonymous reviewers for feedback on an earlier version of this paper. Any remaining errors or shortcomings are my own.

[15] http://russianword.net/en/

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

Cati Brown, Tony Snodgrass, Michael A. Covington, Ruth Herman, and Susan J. Kemper. 2007. Measuring propositional idea density through part-of-speech tagging. poster presented at Linguistic Society of America Annual Meeting, Anaheim, California, January.

Cati Brown, Tony Snodgrass, Susan J. Kemper, Ruth Herman, and Michael A. Covington. 2008. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2):540–545.

Jeanne S. Chall and Edgar Dale. 1995. *Readability revisited: the new Dale-Chall Readability Formula*. Brookline Books.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Kevyn Collins-Thompson and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*, Boston, USA.

Scott A. Crossley, David F. Dufty, Philip M. McCarthy, and Danielle S. McNamara. 2007a. Toward a new readability: A mixed model approach. In Danielle S. McNamara and Greg Trafton, editors, *Proceedings of the 29th annual conference of the Cognitive Science Society*. Cognitive Science Society.

Scott A. Crossley, Max M. Louwerse, Philip M. McCarthy, and Danielle S. McNamara. 2007b. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30.

Scott A. Crossley, Jerry Greenfield, and Danielle S. McNamara, 2008. *Assessing text readability using cognitively based indices*, pages 475–493. Teachers of English to Speakers of Other Languages, Inc. 700 South Washington Street Suite 200, Alexandria, VA 22314.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational research bulletin; organ of the College of Education*, 27(1):11–28.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

William H. DuBay. 2006. *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

P Erbaggio, S Gopalakrishnan, S Hobbs, and H Liu. 2010. Enhancing student engagement through online authentic materials. *International Association for Language Learning Technology*, 42(2).

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.*

Lijun Feng. 2010. *Automatic Readability Assessment*. Ph.D. thesis, City University of New York (CUNY).

Anastasija Vladimirovna Filippova. 2010. *Upravlenie kačestvom učebnyx materialov na osnove analize trudnosti ponimanija učebnyx tekstov [Managing the quality of educational materials on the basis of analyzing the difficulty of understanding educational texts]*. Ph.D. thesis, Ufa State Aviation Technology University.

Thomas Francois and Patrick Watrin. 2011. On the contribution of MWE-based features to a readability formula for French as a foreign language. In *Proceedings of Recent Advances in Natural Language Processing*, pages 441–447.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, pages 95–126.

Alex Gilmore. 2007. Authentic materials and authenticity in foreign language learning. *Language teaching*, 40(02):97–118.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louweerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers*, 36:193–202.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Mark A Hall. 1999. *Correlation-based feature selection for machine learning*. Ph.D. thesis, The University of Waikato.

Julia Hancke, Detmar Meurers, and Sowmya Vajjala. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbay, India.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Associa-*

tion for Computational Linguistics (HLT-NAACL-07), pages 460–467, Rochester, New York.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008a. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications at ACL-08*, Columbus, Ohio.

Michael Heilman, Le Zhao, Juan Pino, and Maxine Eskenazi. 2008b. Retrieval of reading materials for vocabulary and reading practice. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, pages 80–88, Columbus, Ohio.

V. V. Ivanov. 2013. K voprocu o vozmožnosti ispol'zovanija lingvističeskix xarakteristik složnosti teksta pri issledovanii okulomotornoj aktivnosti pri čtenii u podrostkov [toward using linguistic profiles of text complexity for research of oculomotor activity during reading by teenagers]. *Novye issledovanija [New studies]*, 34(1):42–50.

Nikolay Karpov, Julia Baranova, and Fedor Vitugin. 2014. Single-sentence readability prediction in Russian. In *Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST)*.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.

J. P. Kincaid, R. P. Jr. Fishburne, R. L. Rogers, and B. S Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN.

A. Kotlyarov. 2015. Measuring and analyzing comprehension difficulty of texts in contemporary Russian. In *Materials of the annual scientific and practical conference of students and young scientists (with international participation)*, pages 63–65, Kostanay, Kazakhstan.

Nikolaj Konstantinovič Krioni, Aleksej Dmitrievič Nikin, and Anastasija Vladimirovna Filippova. 2008. Avtomatizirovannaja sistema analiza složnosti učebnyx tekstov [automated system for analyzing the complexity of educational texts]. *Vestnik Ufimskogo Gosudarstvennogo Aviacionnogo Texničeskogo Universiteta [Bulletin of the Ufa State Aviation Technical University]*, 11(1):101–107.

O. N. Ljaševskaja and S. A. Šarov. 2009. *Častotnyj slovar' sovremennogo russkogo jazyka (na materialax Nacional'nogo Korpusa Russkogo Jazyka) [Frequency dictionary of Modern Russian (based on the Russian National Corpus)]*. Azbukovnik, Moscow.

Ja. A. Mikk. 1974. Metodika razrabotki formul čitabel'nosti [methods for developing readability formulas]. *Sovetskaja pedagogika i škola IX*, page 273.

I. Ju. Mizernov and L. A. Graščenko. 2015. Analiz metodov ocenki složnosti teksta [analysis of methods for evaluating text complexity]. *Novye informacionnye texnologii v avtomatizirovannyx sistemax [New information technologies in automated systems]*, 18:572–581.

Keith Morrow. 1977. Authentic texts in ESP. *English for specific purposes*, pages 13–16.

Aleksej Dmitrievič Nikin, Nikolaj Konstantinovič Krioni, and Anastasija Vladimirovna Filippova. 2007. Informacionnaja sistema analiza učebnogo teksta [information system for analyzing educational texts]. In *Trudy XIV Vserossijskoj naučno-metodičskoj konferencii Telematika [Proceedings of the XIV pan-Russian scientific-methodological conference Telematika]*, pages 463–465.

Irina Vladimirovna Oborneva. 2005. Matematičeskaja model' ocenki učebnyx tekstov [mathematical model of evaluation of scholastic texts]. In *Informacionnye texnologii v obrazovanii: XV Meždunarodaja konferencija-vystavka [Information technology in education: XV international conference-exhibit*.

Irina Vladimirovna Oborneva. 2006a. Avtomatizacija ocenki kačestva vosprijatija teksta [automation of evaluating the quality of text comprehension]. No longer available on internet.

Irina Vladimirovna Oborneva. 2006b. *Avtomatizirovannaja ocenka složnosti učebnyx tekstov na osnove statističeskix parametrov [Automatic evaluation of the complexity of educational texts on the basis of statistical parameters]*. Ph.D. thesis.

Svetlana Vladimirovna Okladnikova. 2010. Model' kompleksnoj ocenki čitabel'nosti testovyx materialov na etape razrabotki [a model of multidimensional evaluation of the readability of test materials at the development stage]. *Prikaspijskij žurnal: upravlenie i vysokie texnologii*, 3:63–71.

Timothy Osborne, Michael Putnam, and Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15(4):354–396.

Niels Ott. 2009. Information retrieval for language learning: An exploration of text difficulty measures. ISCL master's thesis, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen, Germany.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23:86–106.

I. Ju. Petrova and S. V. Okladnikova. 2009. Metodika rasčeta bazovyx pokazatelej čitabel'nosti testovyx materialov na osnove ekspertnyx ocenok [method of calculating basic indicators of readability of test materials on the basis of expert evaluations]. *Prekaspijskij žurnal: upravlenie i vysokie texnologii*, page 85.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2015. A readable read: Automatic assessment of language learning materials based on linguistic complexity. In *Proceedings of CICLING 2015- Research in Computing Science Journal Issue (to appear)*.

Mikael Roll, Johan Frid, and Merie Horne. 2007. Measuring syntactic complexity in spontaneous spoken Swedish. *Language and Speech*, 50(2):227–245.

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 523–530, Ann Arbor, Michigan.

Serge Sharoff, Svitlana Kurella, and Anthony Hartley. 2008. Seeking needles in the web's haystack: Finding texts suitable for language learners. In *Proceedings of the 8th Teaching and Language Corpora Conference (TaLC-8)*, Lisbon, Portugal.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*, pages 574–576. ACM.

Jurij Francevič Špakovskij, 2003. *Formuly čitabel'nosti kak metod ocenki kačestva knigi [Formulae of readability as a method of evaluating the quality of a book]*, pages 39–48. Ukrainska akademija drukarstva, Lviv'.

Jurij Francevič Špakovskij. 2008. Razrabotka količestvennoj metodiki ocenki trudnosti vosprijatija učebnyx tekstov dl'a vyssšej školy [development of quantitative methods of evaluating the difficulty of comprehension of educational texts for high school]. *Naučno-texničeskij vestnik [Instructional-technology bulletin]*, pages 110–117.

A. Jackson Stenner. 1996. Measuring reading comprehension with the lexile framework. In *Fourth North American Conference on Adolescent/Adult Literacy*.

A. N. Tixonov. 2002. *Morfemno-orfografičeskij slovar': okolo 100 000 slov [Morpho-orthographic dictionary: approx 100 000 words]*. AST/Astrel', Moskva.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In Joel Tetreault, Jill Burstein, and Claudial Leacock, editors, *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, Montréal, Canada, June. Association for Computational Linguistics.

Sowmya Vajjala. 2015. *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Ph.D. thesis, University of Tübingen.

Tim Vor der Brück and Sven Hartrumpf. 2007. A semantically oriented readability checker for German. In Zygmunt Vetulani, editor, *Proceedings of the 3rd Language & Technology Conference*, pages 270–274, Poznań, Poland. Wydawnictwo Poznańskie.

Tim Vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep syntactic and semantic indicators. *Informatica*, 32(4):429–435.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics. Corpus available: `http://ilexir.co.uk/applications/clc-fce-dataset`.