Grammar Error Correction in Morphologically Rich Languages: The Case of Russian

Alla Rozovskaya
Queens College, City University of New York
arozovskaya@qc.cuny.edu

Dan Roth
University of Pennsylvania
danroth@seas.upenn.edu

Abstract

Until now, most of the research in grammar error correction focused on English, and the problem has hardly been explored for other languages. We address the task of correcting writing mistakes in morphologically rich languages, with a focus on Russian. We present a corrected and error-tagged corpus of Russian learner writing and develop models that make use of existing state-of-the-art methods that have been well studied for English. Although impressive results have recently been achieved for grammar error correction of non-native English writing, these results are limited to domains where plentiful training data are available. Because annotation is extremely costly, these approaches are not suitable for the majority of domains and languages. We thus focus on methods that use "minimal supervision"; that is, those that do not rely on large amounts of annotated training data, and show how existing minimal-supervision approaches extend to a highly inflectional language such as Russian. The results demonstrate that these methods are particularly useful for correcting mistakes in grammatical phenomena that involve rich morphology.

1 Introduction

This paper addresses the task of correcting errors in text. Most of the research in the area of grammar error correction (GEC) focused on correcting mistakes made by English language learners. One standard approach to dealing with these errors, which proved highly successful in text correction competitions (Dale and Kilgarriff, 2011; Dale et al., 2012; Ng et al., 2013, 2014; Rozovskaya et al., 2017), makes use of a *machine-learning classifier paradigm* and is based on the methodology for correcting context-sensitive spelling mistakes (Golding and Roth, 1996, 1999; Banko and Brill, 2001). In this approach, classifiers are trained for a particular mistake type: for example, preposition, article, or noun number (Tetreault et al., 2010; Gamon, 2010; Rozovskaya and Roth, 2010c,b; Dahlmeier and Ng, 2012). Originally, classifiers were trained on native English data. As several annotated learner datasets became available, models were also trained on annotated learner data.

More recently, the statistical machine translation (MT) methods, including neural MT, have gained considerable popularity thanks to the availability of large annotated corpora of learner writing (e.g., Yuan and Briscoe, 2016; Junczys-Dowmunt and Grundkiewicz, 2016; Chollampatt and Ng, 2018). Classification methods work very well on well-defined types of errors, whereas MT is good at correcting interacting and complex types of mistakes, which makes these approaches complementary in some respects (Rozovskaya and Roth, 2016).

Thanks to the availability of large (in-domain) datasets, substantial gains in performance have been made in English grammar correction. Unfortunately, research on other languages has been scarce. Previous work includes efforts to create annotated learner corpora for Arabic (Zaghouani et al., 2014), Japanese (Mizumoto et al., 2011), and Chinese (Yu et al., 2014), and shared tasks on Arabic (Mohit et al., 2014; Rozovskaya et al., 2015) and Chinese error detection (Lee et al., 2016; Rao et al., 2017). However, building robust models in other languages has been a challenge, since an approach that relies on heavy supervision is not viable across languages, genres, and learner backgrounds. Moreover, for languages that are complex morphologically, we may need more data to address the lexical sparsity.

Transactions of the Association for Computational Linguistics, vol. 7, pp. 1–17, 2019. Action Editor: Jianfeng Gao.
Submission batch: 4/2018; Revision batch: 8/2018; Published 3/2019.

This work focuses on Russian, a highly inflectional language from the Slavic group. Russian has over 260M speakers, for 47% of whom Russian is not their native language.[1] We corrected and error-tagged over 200K words of non-native Russian texts. We use this dataset to build several grammar correction systems that draw on and extend the methods that showed state-of-the-art performance on English grammar correction. Because the size of our annotation is limited, compared with what is used for English, one of the goals of our work is to quantify the effect of having limited annotation on existing approaches. We evaluate both the MT paradigm, which requires large amounts of annotated learner data, and the classification approaches that can work with any amount of supervision.

Overall, the results obtained for Russian are much lower than those reported for English. We further find that the minimal-supervision classification methods that can combine large amounts of native data with a small annotated learner sample give the best results on a language with rich morphology and with limited annotation. The system that uses classifiers with minimal supervision achieves an $F_{0.5}$ score of 21.0,[2] whereas the MT system trained on the same data achieves a score of only 10.6.

This paper makes the following contributions: (1) We describe an error classification schema for Russian learner errors, and present an error-tagged Russian learner corpus. The dataset is available for research[3] and can serve as a benchmark dataset for Russian, which should facilitate progress on grammar correction research, especially for languages other than English. (2) We present an analysis of the annotated data, in terms of error rates, error distributions by learner type (foreign and heritage), as well as comparison to learner corpora in other languages. (3) We extend state-of-the-art grammar correction methods to a morphologically rich language and, in particular, identify classifiers needed to address mistakes that are specific to these languages. (4) We demonstrate that the classification framework with minimal supervision is particularly useful for morphologically rich languages; they can benefit from large amounts of native data, due to a large variability of word forms, and small amounts of annotation provide good estimates of typical learner errors. (5) We present an error analysis that provides further insight into the behavior of the models on a morphologically rich language.

Section 2 presents related work. Section 3 describes the corpus. Experiments are described in Section 4, and the results are presented in Section 5. We present an error analysis in Section 6 and conclude in Section 7.

2 Background and Related Work

We first discuss related work in text correction on languages other than English. We then introduce the two frameworks for grammar correction (evaluated primarily on English learner datasets) and discuss the "minimal supervision" approach.

2.1 Grammar Correction in Other Languages

The two most prominent attempts at grammar error correction in other languages are shared tasks on Arabic and Chinese text correction. In Arabic, a large-scale corpus (2M words) was collected and annotated as part of the QALB project (Zaghouani et al., 2014). The corpus is fairly diverse: it contains machine translation outputs, news commentaries, and essays authored by native speakers and learners of Arabic. The learner portion of the corpus contains 90K words (Rozovskaya et al., 2015), including 43K words for training. This corpus was used in two editions of the QALB shared task (Mohit et al., 2014; Rozovskaya et al., 2015). There have also been three shared tasks on Chinese grammatical error diagnosis (Lee et al., 2016; Rao et al., 2017, 2018). A corpus of learner Chinese used in the competition includes 4K units for training (each unit consists of one to five sentences).

Mizumoto et al. (2011) present an attempt to extract a Japanese learners' corpus from the revision log of a language learning Web site (Lang-8). They collected 900K sentences produced by learners of Japanese and implemented a character-based MT approach to correct the

[1] https://en.wikipedia.org/wiki/Russian_language.

[2] This is a standard metric used in grammar correction since the CoNLL shared tasks. Because precision is more important than recall in grammar correction, it is weighed twice as high, and is denoted as $F_{0.5}$. Other metrics have been proposed recently (Felice and Briscoe, 2015; Napoles et al., 2015; Choshen and Abend, 2018a).

[3] https://github.com/arozovskaya/RULEC-GEC.

errors. The English learner data from the Lang-8 Web site is commonly used as parallel data in English grammar correction. One problem with the Lang-8 data is a large number of remaining unannotated errors.

In other languages, attempts at automatic grammar detection and correction have been limited to identifying specific types of misuse (grammar or spelling). Imamura et al. (2012) address the problem of particle error correction for Japanese, and Israel et al. (2013) develop a small corpus of Korean particle errors and build a classifier to perform error detection. De Ilarraza et al. (2008) address errors in postpositions in Basque, and Vincze et al. (2014) study definite and indefinite conjugation usage in Hungarian. Several studies focus on developing spell checkers (Ramasamy et al., 2015; Sorokin et al., 2016; Sorokin, 2017).

There has also been work that focuses on annotating learner corpora and creating error taxonomies that do not build a grammar correction system. Dickinson and Ledbetter (2012) present an annotated learner corpus of Hungarian; Hana et al. (2010) and Rosen et al. (2014) build a learner corpus of Czech; and Abel et al. (2014) present KoKo, a corpus of essays authored by German secondary school students, some of whom are non-native writers. For an overview of learner corpora in other languages, we refer the reader to Rosen et al. (2014).

2.2 Approaches to Text Correction

There are currently two well-studied paradigms that achieve competitive results on the task in English–MT and machine learning classification. In the classification approach, error-specific classifiers are built. Given a **confusion set**, for example {*a, the, zero_article*} for articles, each occurrence of a confusable word is represented as a vector of features derived from a **context window** around it. Classifiers can be trained either on learner or on native data, where each target word occurrence (e.g., *the*) is treated as a positive training example for the corresponding word. Given a text to correct, for each confusable word, the task is to select the most likely candidate from the relevant confusion set. Error-specific classifiers are typically trained for common learner errors—for example, article, preposition, or noun number in English (Izumi et al., 2003; Han et al., 2006; Gamon et al., 2008; De Felice and Pulman, 2008; Tetreault et al., 2010; Gamon, 2010; Rozovskaya and Roth, 2011; Dahlmeier and Ng, 2012).

In the MT approach, the error correction problem is cast as a translation task: namely, translating ungrammatical learner text into well-formed grammatical text, and original learner texts and the corresponding corrected texts act as parallel data. MT systems for grammar correction are trained using 20M–50M words of learner texts to achieve competitive performance. The MT approach has shown state-of-the-art results on the benchmark CoNLL-14 test set in English (Susanto et al., 2014; Junczys-Dowmunt and Grundkiewicz, 2016; Chollampatt and Ng, 2017); it is particularly good at correcting complex error patterns, which is a challenge for the classification methods (Rozovskaya and Roth, 2016). However, phrase-based MT systems do not generalize well beyond the error patterns observed in the training data. Several neural encoder–decoder approaches relying on recurrent neural networks were proposed (Chollampatt et al., 2016; Yuan and Briscoe, 2016; Ji et al., 2017). These initial attempts were not able to reach the performance of the state-of-the-art phrase-based MT systems (Junczys-Dowmunt and Grundkiewicz, 2016), but more recently neural MT approaches have shown competitive results on English grammar correction (Chollampatt and Ng, 2018; Junczys-Dowmunt et al., 2018; Junczys-Dowmunt and Grundkiewicz, 2018).[4] However, neural MT systems tend to require even more supervision. For instance, Junczys-Dowmunt et al. (2018) adopt the methods developed for low-resource machine translation tasks, but they still require parallel corpora in tens of millions of tokens.

Minimal Supervision Framework As we have noted, classifiers can be trained on either native or learner data. Native data are cheap and available in large quantities. But, when training on learner data, the potentially erroneous word can also be used by the model. Because mistakes

[4] Single neural MT systems are still not as good as a phrase-based system (Junczys-Dowmunt and Grundkiewicz, 2018), and the top results are achieved using an ensemble of neural models (Chollampatt and Ng, 2018) or a pipeline of a phrase-based and a neural model enhanced with a spell checker (Junczys-Dowmunt and Grundkiewicz, 2018).

made by non-native speakers are not random (Montrul and Slabakova, 2002; Ionin et al., 2008), using the potentially erroneous word and the correction provides the models with knowledge about learner error patterns. For this reason, models trained on error-annotated data often outperform models trained on larger amounts of native data (Gamon, 2010; Dahlmeier and Ng, 2011). But this approach requires large amounts of annotated learner data (Gamon, 2010). The minimal supervision approach (Rozovskaya and Roth, 2014; Rozovskaya et al., 2017) incorporates the best of both modes: training on native texts to facilitate the possibility of training from large amounts of data without the need for annotation, but using a modest amount of expensive learner data that contains learner error patterns. Importantly, error patterns can be estimated robustly with a small amount of annotation (Rozovskaya et al., 2017). The error patterns can be provided to the model in the form of artificial errors or by changing the model priors. In this work, we use the artificial errors approach; it has been studied extensively for English grammar correction. Several other studies consider the effect of using artificial errors (e.g., Cahill et al., 2013; Felice and Yuan, 2014).

3 Corpus and Annotation

We annotated data from the Russian Learner Corpus of Academic Writing (RULEC, 560K words) (Alsufieva et al., 2012), which consists of essays and papers written in a university setting in the United States by students learning Russian as a foreign language and heritage speakers (those who grew up in the United States but had exposure to Russian at home). This closely mirrors the datasets used for English grammar correction. The corpus contains data from 15 foreign language learners and 13 heritage speakers. RULEC is freely available for research use.[5]

3.1 Russian Grammatical Categories

Russian is a fusional language with free word order, characterized by rich morphology and a high number of inflections. Nouns, adjectives, and certain pronouns are specified for gender, number, and case. Modifiers agree with the head nouns; thus, words in these grammatical categories can have up to 24 different word forms. Verbs are marked for number, gender, and person and agree with the grammatical subject. Other categories for verbs are aspect, tense, and voice. These are typically expressed through morphemes corresponding to functional words in English (shall, will, was, have, had, been, etc.).

3.2 Annotation

Two annotators, native speakers of Russian with a background in linguistics, corrected a subset of RULEC (12,480 sentences, comprising 206K words). One of the annotators is an English as a Second Language instructor and English–Russian translator. The annotation was performed using a tool built for a similar annotation project for English (Rozovskaya and Roth, 2010a). We refer to the resulting corpus as RULEC-GEC.

When selecting sentences to be annotated, we attempted to include a variety of writers from each group (foreign and heritage speakers). The annotated data include 12 foreign and 5 heritage writers. The essays of each writer were sorted alphabetically by the essay file name; the essays for annotation were selected in that order, and the sentences were selected in the order they appear in each essay. We intentionally selected more essays from non-native authors, as we conjectured that these authors would display a greater variety of grammatical errors and higher error rates. Eventually, for each author, a subset of that writer's essays was included, but a different number of annotated essays per author, namely, between 13 and 159 essays per author.

The data were corrected, and each mistake was assigned a type. We developed an *error classification schema* that addresses errors in morphology, syntax, and word usage, and takes into account linguistic properties of the Russian language, by emphasizing those that are most commonly misused. The common phenomena were identified through a pilot annotation, and with the help of sample errors that had been collected with the Russian National Corpus in the process of developing a similar annotation of Russian learner texts. The sample errors were made available to us by the authors (Klyachko et al., 2013). This study resulted in an annotated corpus, available for online search at `http://web-corpora.net/` (Rakhilina et al., 2016).

[5] `https://github.com/arozovskaya/RULEC-GEC`.

Noun:case

Это	зависит	от	*показания/показаний	очевидцев
This	depends	from	$testimony_{\text{gen,*sg/gen,pl}}$	$eyewitness_{\text{gen,pl}}$

'This depends on the testimony of eyewitnesses'

Preposition

Слова	*от/из	прошлых	уроков
$word_{\text{nom,pl}}$	*from**/out of**	$previous_{\text{gen,pl}}$	$lesson_{\text{gen,pl}}$

'Words from previous lessons'

Verb number agreement

Все	новые	здания	*разваливается/разваливаются	
All	$new_{\text{nom,pl}}$	$building_{\text{nom,pl}}$	$*fall_{\text{pres,imperfect,sg}}/fall_{\text{pres,imperfect,pl}}$	$apart$

'All new buildings are falling apart'

Verb gender agreement

Лера	*пробовал/пробовала	флиртовать	с	ним
Valerie	$*try_{\text{past,imperfect,masc}}/try_{\text{past,imperfect,fem}}$	to flirt	with	him

'Valerie tried flirting with him'

Lexical choice

Тогда	люди	стали	*спрашивать/задавать	вопросы
Then	$people_{\text{nom,pl}}$	started	*to inquire/**to ask**	$questions_{\text{acc,pl}}$

'Then people started to ask questions'

Table 1: Examples of common errors in the Russian learner corpus. Incorrect words are marked with an asterisk.

Annotator	Total words	Corrected words	Error rate (%)
Annotator A	77,494	5,315	6.9
Annotator B	128,764	7,732	6.0
Total	206,258	13,047	6.3

Table 2: Statistics for the annotated data in RULEC-GEC.

Error type	Total	%	Errors per 1,000 words
Spelling	2575	21.7	12.5
Noun:case	1560	13.2	7.6
Lexical choice	1451	12.3	7.0
Punctuation	1139	9.6	5.5
Missing word	989	8.4	4.8
Replace	687	5.8	3.3
Extra word	618	5.2	3.0
Adj.:case	428	3.6	2.1
Preposition	364	3.1	1.8
Word form	354	3.0	1.7
Noun:number	286	2.4	1.4
Verb:number/gender	285	2.4	1.4
Verb:aspect	208	1.8	1.0

Table 3: Distribution by error type. Total number of categories is 23. The top 13 are shown. *Replace* includes phenomena not covered by other categories, e.g., additional morphological phenomena, replacing multi-word expressions, and word order.

Our error tagset was developed independently and is smaller than the one in Rakhilina et al. (2016), in order to minimize the annotation burden, while still being able to distinguish among most typical linguistic problems for Russian language learners. We include 23 tags that cover syntactic and morphosyntactic errors, orthography, and lexical errors. Table 1 illustrates some of the common errors, and Table 2 presents annotation statistics. Frequencies for the top 13 errors are shown in Table 3. Note that the top 10 error types account for over 80% of all errors. Not shown are the phenomena that occur less than one error per 1,000 words: *adj:gender*, *verb:voice*, *verb:tense*, *adj:other*, *pronoun*, *adj:number*, *conjunction*, *verb:other*, *noun:gender*, *noun:other*.

3.3 Inter-Annotator Agreement

Because annotation for grammatical errors is extremely variable, as there are often multiple ways of correcting the same mistake (Bryant and Ng, 2015), we compute inter-rater agreement following Rozovskaya and Roth (2010a), where the texts corrected by one annotator were given to the second annotator. Agreement is computed as the percentage of sentences that did not have additional corrections on the second pass. After

Second pass	Error rate (%)	Judged correct (%)
Annotator A	2.40	68.5
Annotator B	0.67	91.5

Table 4: Inter-annotator agreement. *Error rates* based on the corrections on the second pass. *Judged correct* denotes the percentage of sentences that the second rater did not change.

all, our goal is to make the sentence well formed, without enforcing that errors are corrected in the same way. A total of 200 sentences from each annotator were selected and given to the other annotator. Table 4 shows that the error rate of the sentences corrected by annotator A on the second pass was 2.4%, with 68.5% of the sentences remaining unchanged. The sentences corrected by annotator B on the second pass had an error rate of less than 1%, and over 91% of the sentences did not have additional corrections. These agreement numbers are higher than those reported for English, where the percentage of unchanged sentences varied between 37% and 83% (Rozovskaya and Roth, 2010a).

3.4 Comparison to Other Learner Corpora

Error Rates In Table 5, we compare the error rates in RULEC-GEC to those in a learner corpus of Arabic (Zaghouani et al., 2014) and three corpora of learner English: JFLEG (Napoles et al., 2017), FCE (Yannakoudakis et al., 2011), and CoNLL (Ng et al., 2014). The error rates in RULEC are generally lower than in the other learner corpora. The Arabic data have the highest error rate of 28.7%. In the English learner corpora, the error rates range between 6.5% and 25.5%. The error rates are 17.7% (FCE); 18.5–25.5% for JFLEG, annotated independently by four raters; and 10.8–13.6% for CoNLL-test, annotated by two raters. The lowest error rate that is comparable to ours is in CoNLL-train (6.6%). We attribute the differences to the proficiency levels of the RULEC writers, which is fairly advanced. In fact, error rates vary widely by learner group (foreign vs. heritage), as discussed in Section 3.5.

Most Common Errors Table 6 lists the top five most common errors for the three corpora (the Arabic corpus and JFLEG are not annotated for types of errors). In English, lexical choice errors, article, preposition, punctuation, and spelling are

Corpus	Error rate (%)
Russian (RULEC-GEC)	6.3
English (FCE)	17.7
English (CoNLL-test)	10.8–13.6
English (CoNLL-train)	6.6
English (JFLEG)	18.5–25.5
Arabic	28.7

Table 5: Error rates in various learner corpora. The CoNLL and JFLEG have two and four reference annotations, respectively. Numbers shown for each.

the most common mistake types (note that "mechanical errors" in CoNLL group together spelling and punctuation errors). Noun number errors are also common in CoNLL, a corpus produced by learners whose first language is Chinese, whereas these are less common in FCE, produced by learners of diverse linguistic backgrounds. In Russian, spelling, punctuation, and lexical choice are also in the top five.

In RULEC-GEC, the top five error categories are spelling, lexical choice, noun:case, punctuation, and missing word. Overall, spelling, punctuation, and lexical errors are in the top five categories for all of the three corpora. As for grammar-related errors, although article and preposition errors also made it to the top of the list in the English corpora, noun case usage is definitely the most challenging and common phenomenon for Russian learners.

3.5 Foreign vs. Heritage Speakers

We also compare foreign and heritage speakers. The heritage speaker subcorpus includes 42,187 words, and the foreign speaker partition comprises 164,071 words. The error rates are 4.0% and 6.9% for each group; foreign learners make almost twice as many mistakes as heritage speakers. In the foreign group, there is a lot of variation, with five writers exhibiting error rates of 10–13%, two writers whose error rates are below 3%, and five authors having error rates between 5% and 7%. There is not much variation in the heritage group.

The two groups also reveal differences in the error distributions (Table 7): More than 65% of errors in the heritage group are in spelling and punctuation. In fact, 42.4% of errors in the heritage corpus are spelling mistakes vs. 18.6% for foreign speakers. If we consider the number of errors per 1,000 words, we observe that, interestingly,

Top errors (%)		
Russian	**English (FCE)**	**English (CoNLL-test)**
Spell. 21.7	Art. 11.0	Lex. choice 14.2/14.4
Noun:case 13.2	Lex. choice 9.5	Art. 13.9/13.3
Lex. choice 12.3	Prep. 9.0	Mechan. 9.6/14.9
Punc. 9.6	Spell. 8.1	Noun:number 9.0/6.8
Miss. word 8.4	Punc. 8.0	Prep. 8.8/11.7

Table 6: Comparison statistics for Russian and English learner corpora. The CoNLL-test was annotated by two annotators; numbers shown for each.

Foreign			Heritage		
Error	**(%)**	**Errors per 1,000**	**Error**	**(%)**	**Errors per 1,000**
Spell.	18.6	11.7	Spell.	42.4	15.7
Noun:case	14.0	8.8	Punc.	22.9	8.5
Lex. choice	13.3	8.3	Noun:case	7.8	2.9
Miss. word	8.9	5.6	Lex. choice	5.5	2.0
Punc.	7.6	4.8	Miss. word	4.7	1.7
Replace	6.3	3.9	Replace	2.8	1.0
Extra word	5.7	3.5	Extra word	2.4	0.9
Adj:case	3.9	2.4	Adj:case	2.1	0.8
Prep.	3.3	2.1	Word form	2.1	0.8
Word form	3.1	2.0	Noun:number	1.8	0.7
Noun:number	2.6	1.6	Verb agr.	1.6	0.6
Verb agr.	2.5	1.6	Prep.	1.5	0.6

Table 7: Most common errors for foreign and heritage Russian speakers.

heritage speakers make spelling and punctuation errors more frequently (15.7 spelling and 8.5 punctuation errors in the heritage group vs. 11.7 spelling and 4.8 punctuation errors in the foreign group). As for the other grammatical phenomena, although these are all more challenging for the foreign speaker group, the distributions of these phenomena are quite similar. For example, heritage speakers make 2.9 noun case errors per 1,000 words, whereas foreign speakers make 8.8 noun case errors per 1,000 words; for both types of writers, noun case errors are at the top of the list (second most common for the foreign group and third most common for the heritage group).

4 Experiments

The experiments investigate the following:

1. How do the two state-of-the-art methods compare under the conditions that we have for Russian (rich morphology and limited annotations)?
2. What is the performance on individual errors and the overall performance compared with results obtained for English grammar correction?
3. How well do the classifiers within the minimal supervision framework perform in morphologically rich languages, on grammatical phenomena that are common in highly inflectional languages such as Russian, as well as on phenomena that also occur in English?

To answer these questions, the following three approaches are implemented:

- *Learner-trained classifiers*: Error-specific classifiers trained on learner data
- *Minimal-supervision classifiers*: Error-specific classifiers trained on learner and native data with minimal supervision (see Section 2.2)
- *Phrase-based machine translation* system

Data We split the annotated data into training (4,980 sentences, 83,410 words), development (2,500 sentences, 41,163 words), and test (5,000

sentences, 81,693 words). For the native data, we use the Yandex corpus (Borisov and Galinskaya, 2014), a diverse corpus of newswire, fiction, and other genres (18M words). All the data was pre-processed with the the Mystem morphological analyzer (Segalovich, 2003) and a part-of-speech tagger (Schmid, 1995).

4.1 Classifiers

In the classification framework, we develop classifiers for several common grammar errors: preposition, noun case, verb aspect, and verb agreement (split into number and gender). The rationale for selecting these errors is to evaluate the behavior of the classifiers on phenomena that have been well studied in English (e.g., preposition and verb number agreement), as well as those that have not received much attention (verb aspect); or those that are specific to Russian (noun case and gender agreement). For each error type, a special classifier is developed. The features include word n-grams, POS n-grams, lemma n-grams, and morphological properties of the target word and neighboring words. In addition, in line with Rozovskaya and Roth (2016), we include a punctuation module that inserts missing commas, using patterns mined from the Yandex corpus and the RULEC-GEC training data. We now provide more detail on the grammar phenomena considered.

Noun Case Errors Noun case usage is the most common error type after spelling and accounts for 14% of all errors. The Russian case system consists of six cases: Nominative, Genitive, Accusative, Dative, Instrumental, and Locative. The case classifier is thus a six-way classifier, with each class corresponding to one of the cases. The labels are obtained by extracting the case information predicted by the morphological analyzer on original and corrected noun forms. It should be noted that the surface form of the noun may be ambiguous with respect to case. For example, the word яблоко (“apple”) in different contexts can be interpreted as nominative or accusative. In that case, the morphological analyzer will list both analyses, and both of these will be included as gold labels for the word. This is because our task is not to predict the case but the surface form of the noun. About 58% of nouns are unambiguous (have one case-related morphological analysis), 34% have two possible case analyses, and 8% of nouns have three or more analyses.

Error	Confusion set
Noun:case	{Nom., Gen., Acc., Dat., Inst., Loc.}
Verb agr. (num.)	{Singular, Plural}
Verb agr. (gender)	{Fem., Masc., Neutral}
Aspect	{Perfect, Imperfect}
Prep.	{15 prepositions}

Table 8: Confusion sets for the five types of errors.

Number and Gender Verb Agreement Errors Verb agreement functions in a way that is similar in English. In Russian, verbs are specified for number (singular, plural), gender (feminine, masculine, and neutral), and person. Errors in person agreement are rare, and we ignore these.

Preposition Errors Preposition errors are some of the most common errors for learners of English (Leacock et al., 2010), and are also quite common among the Russian learners, accounting for over 3% of all errors (Table 3). In the classification framework, it is common to consider top n most frequent prepositions (Dahlmeier and Ng, 2012; Tetreault et al., 2010). In line with work in English, we consider mistakes that involve the top 15 Russian prepositions.[6]

Verb Aspect The Russian verb system is different from English, and verb aspect errors among Russian learners are quite common. Russian has three tenses—present, past, and future—and each tense can be expressed in *imperfective* or *perfective* aspect. Although there is no direct correspondence between the Russian aspect usage and the English tenses, the aspect can be weakly aligned with the English tense system. Prior research in English showed that these are some of the most difficult mistakes, as verb tense usage is highly semantic rather than grammatical (Lee and Seneff, 2008; Tajiri et al., 2012).

Table 8 lists the confusion sets for each error classifier. In all cases, discriminative learning framework is used with the Averaged Perceptron algorithm (Rizzolo, 2011).

Adding Artificial Errors in the Classifiers Within both the learner-trained and minimally supervised classifiers, we make use of the artificial errors

[6] {в (in, to), на (on, to), с (from), о (about), для (for), к (to), из (from), по (along, on), от (from), у (at), за (for, behind), во (in, to), между (between), до (before), об (about)}.

	Sources					
Label	Nom.	Gen.	Dat.	Acc.	Inst.	Loc.
Nom.	**.9961**	.00214	.0002	.0006	.0002	.0007
Gen.	.01147	**.9775**	.0009	.0043	.0031	.0027
Dat.	.0097	.0105	**.9589**	.0105	0045	.0060
Acc.	.0064	.0056	.0004	**.9837**	.0008	.0031
Inst.	.0163	.0181	.0018	.0113	**.9511**	.0014
Loc.	.0029	.0068	-	.0087	.0017	**.0980**

Table 9: Confusion matrix for noun case errors based on the training and development data from the RULEC-GEC corpus. The left column shows the correct case. Each row shows the author's case choices for that label and $Prob(source|label)$.

approach (Rozovskaya et al., 2017) to simulate learner errors in training. Learner error patterns (or error statistics) are extracted from the annotated learner data. Specifically, given an error type, we collect all source/label pairs from the annotated sample, where both the source and the label belong to the confusion set, and generate a confusion matrix, where each cell represents $Prob(\text{source=s}|\text{label=l})$.

Table 9 shows a confusion matrix for noun case errors based on error statistics collected from the training and development data in RULEC-GEC. The values in the confusion matrix are used to generate noun form errors in the training data. For instance, according to the table, given a noun that needs to be in the genitive case, a learner is four times more likely to use the nominative case instead of the locative case. We use this table both to introduce artificial errors in native training data and to increase the error rates in the learner data by adding artificial mistakes to naturally occurring errors. Adding artificial errors when training on learner data is also useful, as increasing the error rates improves the recall of the system. In both cases, the generated errors are added, so that the relative frequencies of different confusions are preserved (e.g., nominative is four times more likely than locative to be used in place of genitive), and the error rates can be varied (higher error rates will improve the recall of the system at the expense of precision).

4.2 The MT System

One advantage of the MT approach is that error types need not be formulated explicitly. We build a phrase-based MT system that follows the implementation in Susanto et al. (2014). Our MT system is trained using Moses (Koehn et al., 2007). The phrase table is trained on the training partition of RULEC-GEC. We use two 4-gram language models—one is trained on the Yandex corpus, and the other one is trained on the corrected side of the RULEC-GEC training data. Both are trained with KenLM (Heafield et al., 2013). Tuning is done on the development dataset with MERT (Och, 2003). We use BLEU (Papineni et al., 2002) as the tuning metric.

We note that several neural MT systems have been proposed recently (see Section 2). Because we only have a small amount of parallel data, we adopt the phrase-based MT, as it is known that neural MT systems have a steeper learning curve with respect to the amount of training data, resulting in worse quality in low-resource settings (Koehn and Knowles, 2017). We also note that Junczys-Dowmunt and Grundkiewicz (2016) present a stronger SMT system for English grammar correction. Their best result that is due to adding dense and sparse features is an improvement of 3 to 4 points over the baseline system (they also rely on much large tuning sets, as required for sparse features). The baseline system is essentially the same as that of Susanto et al. (2014). Because our MT result is so much lower than the classification system, we do not expect that adding sparse and dense features will close that gap.

5 Results

We start by comparing performance on individual errors; then the overall performance of the best classification systems and the MT system is compared.

Classifier Performance On Individual Errors First, we wish to assess the contribution of the minimal-supervision approach compared with training on the learner data for a language with rich

Error	Training data	Performance		
		P	R	$F_{0.5}$
Case	(1) Learner	17.9	32.8	19.7
	(2) Learner+native	34.5	36.1	**34.8**
Number agr.	(1) Learner	56.7	7.7	24.9
	(2) Learner+native	49.1	16.6	**35.3**
Gender agr.	(1) Learner	48.5	7.1	22.4
	(2) Learner+native	67.9	16.0	**41.2**
Aspect	(1) Learner	21.6	2.5	8.6
	(2) Learner+native	21.5	9.1	**16.9**
Prep.	(1) Learner	31.9	3.8	12.9
	(2) Learner+native	56.1	24.9	**44.8**

Table 10: Comparison of classifiers trained on (1) learner data and (2) learner + native data, using the minimal supervision framework.

morphology. To this end, two types of classifiers are compared: learner-trained (trained on learner data) and minimal-supervision (trained on native data with artificial errors based on error statistics extracted from the learner data; Section 2). The classifiers are tuned on the development partition—that is, the error rates that determine at which rate artificial errors injected into the training data are optimized on the development data. Performance results on the test data are for models trained on the training+development data (learner-trained models). Similarly, the minimal supervision classifiers use error statistics extracted from training+development.

Table 10 shows performance for the five types of errors. For all errors, minimal-supervision models outperform the learner-trained models substantially, by 8 to 32 $F_{0.5}$ points. This is because the amount of annotation that we have is really too small to estimate all parameters, but it is sufficient to provide error estimates in the minimal supervision framework. In addition, the punctuation module achieves an $F_{0.5}$ score of 30.5 (precision of 47.4 and recall of 12.6).

Classifiers vs. MT So far, we have evaluated performance of the classifiers with respect to individual errors. Table 11 shows the performance of the three systems on the entire dataset and evaluates with respect to all errors in the data. The results show that when annotation is scarce, MT performs poorly. This result is consistent with findings for English, showing that MT systems outperform classifiers only when the parallel corpus is large (30–40M words) (Rozovskaya and Roth, 2016) but lag behind even when over 1M tokens are available.

System	Training data	P	R	$F_{0.5}$
Classifiers (learner)	Learner	22.6	4.8	12.9
Classifiers (minimal sup.)	Learner+native	38.0	7.5	**21.0**
MT	Learner+native	30.6	2.9	10.6

Table 11: Performance of the three systems.

We combine the MT system and the minimally supervised classifiers following Rozovskaya and Roth (2016). Because MT systems are not restricted for error type, the misuse they correct is typically more diverse (see also Section 6). The $F_{0.5}$ score thus improves by 2 points, to 23.8, for the combined system, due to a slightly better recall (10.2). However, the precision drops from 38.0 to 35.8, since the MT system has a lower precision than the classifiers.

6 Discussion and Error Analysis

The current state of the art in English grammar correction on the widely used benchmark CoNLL test is 50.27 for a single system (Junczys-Dowmunt and Grundkiewicz, 2018). System combination, model ensembles, and adding a spell checker boost these numbers by 4 to 6 points (Chollampatt and Ng, 2018; Junczys-Dowmunt and Grundkiewicz, 2018). These models are trained on the CoNLL training data and additional learner data (about 30M words). An MT system trained on CoNLL data (1.2M words) obtains an $F_{0.5}$ score of 28.25 (Rozovskaya and Roth, 2016). Although these MT systems differ in how they are trained, these numbers should give an idea of the effect the amount of parallel data has on the performance.

Error type	Before			After		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
Case	34.5	36.1	34.8	**38.2**	36.1	**37.8**
Number agr.	49.1	16.6	35.3	**56.7**	35.3	**38.2**
Gender agr.	67.9	16.0	41.2	67.9	16.0	41.2
Prep.	56.1	24.9	44.8	**72.2**	24.9	**52.3**
Aspect	21.5	9.1	16.9	**30.0**	9.1	**20.6**

Table 12: Performance of minimal-supervision classifiers before and after false positive analysis.

A minimal-supervision classification system that uses CoNLL data obtains an $F_{0.5}$ score of 36.26 (Rozovskaya and Roth, 2016). In contrast, the classification system for Russian obtains a much lower score of 21.0. This may be due to a larger variety of grammatical phenomena in Russian, lower error rates, and a high proportion of spelling errors (especially among heritage speakers), which we currently do not specifically target. Note also that the CoNLL-2014 results are based on two gold references for each sentence, while we evaluate with respect to one, and having more reference annotations improves performance (Bryant and Ng, 2015; Sakaguchi et al., 2016; Choshen and Abend, 2018b).[7] It should also be noted that the gap between the MT system and the classification system when both are trained with limited supervision is larger for Russian (10.6 vs. 20.5) than for English (28.25 vs. 36.26). This indicates that the MT system suffers more than classifiers, when the amount of supervision is particularly small, while the morphological complexity of the language is higher.

Considering Arabic and Chinese, where the training data is also limited, the results are also much lower than in English. In Arabic, where the supervised learner data includes 43K words, the best reported F-score is 27.32 (Rozovskaya et al., 2015).[8] In Chinese, the supervised dataset size is about 50K sentences, and the highest reported scores are 26.93 for detection (Rao et al., 2017) and 17.23 for error correction (Rao et al., 2018), respectively. These results confirm that the approaches that rely on large amounts of supervision do not carry over to low-resource settings. It is thus desirable to develop approaches that can be robust with a small amount of supervision, especially when applied to languages that are morphologically more complex than English.

6.1 Error Analysis

To understand the challenges of grammar correction in a morphologically rich language such as Russian, we perform error analysis of the MT system and the classification system that uses minimal supervision. The nature of grammar correction is such that multiple different corrections are often acceptable (Ng et al., 2014). Furthermore, annotators often disagree on what constitutes a mistake, and some gold errors missed by a system may be considered as acceptable usage by another rater. Thus, when a system is compared against the gold truth produced by just one annotator, performance is understated. In fact, the F-score of a system increases with the number of per-sentence annotations (Bryant and Ng, 2015).

Classifiers: False Positives We start by analyzing the cases where the system flagged an error that was not marked in the gold annotation. False positive cases were manually annotated by one of the annotators and acceptable predictions were identified. As expected, because of the variability in the annotators' judgments and possibility of multiple acceptable options, there are false positives that actually should be true positives. We re-evaluate the performance of the classifiers based on the error analysis in Table 12.

For all error types, except gender agreement (which has a high precision of 67.9%), precision improvements range between 4 points and 16 points. The highest improvement is observed for preposition errors: about 48% of false positives are in fact acceptable suggestions. This improvement mirrors the results in English (precision improves from 30% to 70% [Rozovskaya et al., 2017]) and

[7] There is ongoing research on the question of the most appropriate evaluation metric and gold references for grammatical error correction. See Sakaguchi et al. (2016), Choshen and Abend (2018b), and Choshen and Abend (2018c).

[8] This result is based on performance that does not take into account some trivial Arabic-specific normalization corrections.

В этих местах мало *перспектива/перспектив
In these places few $*prospects_{\text{pl,\textbf{nom}}}/prospects_{\text{pl,\textbf{gen}}}$
'There are few prospects in these places'

Example 1: **Case error on a noun following the adverbial "few".**

Он обеспечивает *клиентов/клиентам доступ к информации
It supplies $*clients_{\text{pl,\textbf{gen}}}/clients_{\text{pl,\textbf{dat}}}$ $access_{\text{sg,acc}}$ towards $information_{\text{sg,gen}}$
'It provides clients with access to information'

Example 2: **Case error on a noun governed by the verb "provides".**

На станции *использует/используют разные приборы
At station $use_{\text{3rd-person,*\textbf{sg}/\text{3rd-person,\textbf{pl}}}}$ various tools
'At the station (they) use various tools'

Example 3: **Agreement error on a verb without an explicit subject.**

Она готова была *давать/дать мне все , что нужно
She ready was $*give_{\text{inf,\textbf{imperfect}}}/give_{\text{inf,\textbf{perfect}}}$ to me everything , that necessary
'She was ready to give me everything that is necessary'

Example 4: **Aspect error on a verb that requires wider context beyond sentence.**

can be explained by the fact that preposition usage is highly variable (i.e., many contexts license multiple prepositions [Tetreault and Chodorow, 2008]).

Classifiers: Errors Missed by the System
Although the precision of the classifiers is generally quite good, the recall is much lower, ranging between 36.1% and 24.9% for noun case and preposition errors to 16% for agreement errors and 9.1% for verb aspect errors.

Among the languages studied in the grammar correction research, noun case errors are unique to Russian.[9] But because the appropriate case choice depends on the word governing the noun, one can view case declension to be similar to subject-verb agreement. However, case errors are arguably more challenging because the target noun may be governed by a verb, a preposition, another noun, or even by an adverbial; thus, there is a higher level of ambiguity when identifying the dependency as well as determining the appropriate case. A morphologically rich language such as Russian uses case to express relations that are commonly conveyed by prepositions in English; as a result, verbs that are followed by a direct object and a prepositional object in English appear with two noun phrases, whose relationship to the verb is expressed through appropriate cases. Examples (1) and (2) illustrate two case errors, where the first noun is governed by an adverbial, and the other noun is governed by a verb. An additional challenge is that prepositions and verbs can also license multiple cases. For example, the prepositions на and в can denote location, when followed by a noun in locative case, as well as direction when followed by a noun in dative case.

Analysis of the missed verb agreement errors reveals several challenges; some of these are specific to morphologically rich languages. The main challenge here is identifying the subject of the target verb. Thus, errors on verbs that are located far from the subject head are typically not handled well in both Russian and English; in the Russian corpus, these account for 20% of all missed errors. Because the system currently does not use a parser, we anticipate that adding a parser will improve performance. However, because of Russian's free word order, there are more options for the location of the subject. It is also not uncommon for a subject to be placed after the verb, and 19% of errors that are currently missed occur when the subject is located after the verb. Finally, about 6% of missed errors occur on verbs that have no explicit subject, as in Example (3). In such cases the verb takes the form of third person singular masculine or third person plural.

Compared with other errors, aspect errors exhibit the lowest performance. Appropriate aspect form may require understanding the context around the verb, often beyond the sentence boundaries. Example (4) illustrates an error where, without looking at the wider context, both perfect and imperfect forms are possible. Some

[9] Case errors have certainly been considered in studies that aim at annotating learner corpora, including Czech (Hana et al., 2010) and German (Abel et al., 2014).

verb aspect errors are similar to verb tense errors in English. Studies in English also reported poor performance, a precision of 20% corresponding to a recall of about 20% on verb aspect errors (Tajiri et al., 2012). Our expectation is that with richer context representation, such as identification of temporal relations, one can do better. Some verbs are also ambiguous with respect to aspect; for example, проводить can be translated as "carry out" (imperfective), and "accompany" (perfective).

The MT System Because the output of the MT system does not specify the correction type, our annotator manually analyzed the true positives of the system and classified these for type. The most common true positive corrections of the MT system fall into the following categories: spelling (40%), missing comma (36%), noun:case (13%), and lexical (7%).

We also analyze the false positives. About 15% of the false positives are in fact true positives. As a result, the precision and the F-score of the MT system improve from 30.6 to 41.0 and from 10.6 to 11.4, respectively. Even though the current MT system performs poorly, the analysis supports the findings in English that MT systems correct a more diverse set of errors and, if trained with sufficient supervision, should complement a classification system well.

7 Conclusion

We address the task of correcting writing mistakes in Russian, a morphologically rich language. We correct and error-tag a corpus of Russian learner data. The release of this corpus should facilitate research efforts in grammar correction for languages other than English that do not have many resources available to them. Experiments on that corpus demonstrate that the MT approach performs poorly due to lack of annotated data. The MT system is outperformed substantially by a minimally supervised machine learning classification approach.

Acknowledgments

The authors thank Olesya Kisselev for her help with obtaining the RULEC corpus, and Elmira Mustakimova for sharing the error categories developed at the Russian National Corpus. The authors thank Mark Sammons and the anonymous reviewers for their comments. This work was partially supported by contract HR0011-15-2-0025 with the US Defense Advanced Research Projects Agency (DARPA). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the US Government.

References

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egol Stemle. 2014. KoKo: An L1 learner corpus for German. In *Proceedings of LREC*, pages 2414–2421.

Anna Alsufieva, Olesya Kisselev, and Sandra Freels. 2012. Results 2012: Using flagship data to develop a Russian learner corpus of academic writing. *Russian Language Journal*, 62:79–105.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of ACL*, pages 26–33.

Alexey Borisov and Irina Galinskaya. 2014. Yandex school of data analysis Russian-English machine translation system for WMT14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 66–70.

Christopher Bryant and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of ACL*, pages 697–707.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust systems for preposition error correction using Wikipedia revisions. In *Proceedings of NAACL-HLT*, pages 507–517.

Shamil Chollampatt and Hwee Tou Ng. 2017. Connecting the dots: Towards human-level grammatical error correction. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 327–333.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI*, pages 1–8.

Shamil Chollampatt, Kaveh Taghipour, and Hwee Tou Ng. 2016. Neural network translation models for grammatical error correction. In *Proceedings of IJCAI*, pages 2768–2774.

Leshem Choshen and Omri Abend. 2018a. Automatic metric validation for grammatical error correction. In *Proceedings of ACL*.

Leshem Choshen and Omri Abend. 2018b. Inherent biases in reference-based evaluation for grammatical error correction and text simplification. In *Proceedings of ACL*.

Leshem Choshen and Omri Abend. 2018c. Reference-less measure of faithfulness for grammatical error correction. In *Proceedings of NAACL-HLT*.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Grammatical error correction with alternating structure optimization. In *Proceedings of ACL*, pages 915–923.

Daniel Dahlmeier and Hwee Tou Ng. 2012. A beam-search decoder for grammatical error correction. In *Proceedings of EMNLP-CoNLL*, pages 568–587.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. A report on the preposition and determiner error correction shared task. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 54–62.

Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249.

Rachele De Felice and Stephen G. Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of COLING*, pages 169–176.

Markus Dickinson and Scott Ledbetter. 2012. Annotating errors in a Hungarian learner corpus. In *Proceedings of LREC*.

Mariano Felice and Ted Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of NAACL-HLT*, pages 578–587.

Mariano Felice and Zheng Yuan. 2014. Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at EACL*.

Michael Gamon. 2010. Using mostly native data to correct errors in learners' writing. In *Proceedings of NAACL-HLT*, pages 163–171.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexander Klementiev, William Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of IJCNLP*.

Andrew R. Golding and Dan Roth. 1996. Applying Winnow to context-sensitive spelling correction. In *Proceedings of ICML*.

Andrew R. Golding and Dan Roth. 1999. A Winnow based approach to context-sensitive spelling correction. *Machine Learning*, 34(1-3): 107–130.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Journal of Natural Language Engineering*, 12(2): 115–129.

Jirka Hana, Alexandr Rosen, Svatava Škodová, and Barbora Štindlová. 2010. Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 11–19.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL*, pages 690–696.

Arantza Díaz de Ilarraza, Koldo Gojenola, and Maite Oronoz. 2008. Detecting erroneous uses of complex postpositions in an agglutinative language. In *Proceedings of COLING*, pages 31–34.

Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. 2012. Grammar error correction using pseudo-error sentences and domain adaptation. In *Proceedings of ACL*, pages 388–392.

Tania Ionin, Maria Luisa Zubizarreta, and Salvador Bautista Maldonado. 2008. Sources of linguistic knowledge in the second language acquisition of English articles. *Lingua*, 118:554–576.

Ross Israel, Markus Dickinson, and Sun-Hee Lee. 2013. Detecting and correcting learner Korean particle omission errors. In *Proceedings of IJCNLP*, pages 1419–1427.

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. 2003. Automatic error detection in the Japanese learners' English spoken data. In *Proceedings of ACL*, pages 145–148.

Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao. 2017. A nested attention neural hybrid model for grammatical error correction. In *Proceedings of ACL*, pages 753–762.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proceedings of EMNLP*, pages 1546–1556.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. Near human-level performance in grammatical error correction with hybrid machine translation. In *Proceedings of NAACL-HLT*, pages 284–290.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of NAACL-HLT*, pages 595–606.

Elena Klyachko, Timofey Arkhangelskiy, Olesya Kisselev, and Ekaterina Rakhilina. 2013. Automatic error detection in Russian learner language. In *Proceedings of the First Workshop on Corpus Analysis with Noise in the Signal (CANS)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*, pages 177–180.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool Publishers.

John Lee and Stephanie Seneff. 2008. An analysis of grammatical errors in non-native speech in English. In *Proceedings of the 2008 Spoken Language Technology Workshop*, pages 89–92.

Lung-Hao Lee, Gaoqi Rao, Liang-Chih Yu, Endong Xun, Baolin Zhang, and Li-Ping Chang. 2016. Overview of NLP-TEA 2016 shared task for Chinese grammatical error diagnosis. In *Proceedings of the Third Workshop on Natural Language Processing Techniques for Educational Applications*, pages 1–6.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of IJCNLP*, pages 147–155.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47.

Silvina Montrul and Roumyana Slabakova. 2002. Acquiring morphosyntactic and semantic properties of preterite and imperfect tenses in L2 Spanish. In Perez-Laroux A-T and Liceras J (eds). *The Acquisition of Spanish Morphosyntax: The L1-L2 Connection*. Dordrecht: Kluwer, pages 113–149.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of ACL*, pages 588–593.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of EACL*, pages 229–234.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of CoNLL: Shared Task*, pages 1–14.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of CoNLL: Shared Task*, pages 1–12.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Ekaterina Rakhilina, Anastasia Vyrenkova, Elmira Mustakimova, Alina Ladygina, and Ivan Smirnov. 2016. Building a learner corpus for Russian. In *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*.

Loganathan Ramasamy, Alexandr Rosen, and Pavel Stranák. 2015. Improvements to korektor: A case study with native and non-native Czech. In *Proceedings of Slovenskočeský NLP Workshop*.

Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In *Proceedings of the Fifth Workshop on Natural Language Processing Techniques for Educational Applications*, pages 42–51.

Gaoqi Rao, Baolin Zhang, and Endong Xun. 2017. IJCNLP-2017 task 1: Chinese grammatical error diagnosis. In *Proceedings of IJCNLP*, pages 1–8.

Nicholas Rizzolo. 2011. *Learning Based Programming*. PhD thesis. University of Illinois, Urbana: Champaign.

Alexandr Rosen, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2014. Evaluating and automating the annotation of a learner corpus. In *Proceedings of LREC*, pages 65–92.

Alla Rozovskaya, Houda Bouamor, Wajdi Zaghouani, Ossama Obeid, Nizar Habash, and Behrang Mohit. 2015. The second QALB shared task on automatic text correction for Arabic. In *Proceedings of the ACL Workshop on Arabic Natural Language Processing*, pages 26–35.

Alla Rozovskaya and Dan Roth. 2010a. Annotating ESL errors: Challenges and rewards. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36.

Alla Rozovskaya and Dan Roth. 2010b. Generating confusion sets for context-sensitive error correction. In *Proceedings of EMNLP*, pages 961–970.

Alla Rozovskaya and Dan Roth. 2010c. Training paradigms for correcting errors in grammar and usage. In *Proceedings of NAACL*, pages 154–162.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of ACL*, pages 924–933.

Alla Rozovskaya and Dan Roth. 2014. Building a state-of-the-art grammatical error correction system. In *Transactions of ACL*, pages 419–434.

Alla Rozovskaya and Dan Roth. 2016. Grammatical error correction: Machine translation and classifiers. In *Proceedings of ACL*, pages 2205–2215.

Alla Rozovskaya, Dan Roth, and Mark Sammons. 2017. Adapting to learner errors with minimal supervision. *Computational Linguistics*, 43(4):723–760.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. In *Transactions of ACL*, pages 169–182.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*, pages 47–50.

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

Alexey Sorokin. 2017. Spelling correction for morphologically rich language: A case study of Russian. In *Proceedings of the Sixth Workshop on Balto-Slavic Natural Language Processing*, pages 45–53.

Alexey Sorokin, Alexey Baytin, Irina Galinskaya, Elena Rykunova, and Tatiana Shavrina. 2016. SpellRuEval: The first competition on automatic spelling correction for Russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016"*.

Raymond Hendy Susanto, Peter Phandi, and Hwee Tou Ng. 2014. System combination for grammatical error correction. In *Proceedings of EMNLP*, pages 951–962.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of ACL: Short Papers*, pages 198–202.

Joel Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the COLING Workshop on Human Judgements in Computational Linguistics*, pages 24–32.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. 2010. Using parse features for preposition selection and error detection. In *Proceedings of ACL*, pages 353–358.

Veronika Vincze, János Zsibrita, Péter Durst, and Martina Katalin Szabó. 2014. Automatic error detection concerning the definite and indefinite conjugation in the hunlearner corpus. In *Proceedings of LREC*, pages 26–31.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of ACL*, pages 180–189.

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang. 2014. Overview of grammatical error diagnosis for learning Chinese as a foreign language. In *Proceedings of the Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA)*.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of NAACL-HLT*, pages 380–386.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale arabic error annotation: Guidelines and framework. In *Proceedings of LREC*.