# Multiple Admissibility in Language Learning: Judging Grammaticality Using Unlabeled Data

**Anisia Katinskaia**, **Sardana Ivanova**, **Roman Yangarber**
University of Helsinki, Department of Computer Science, Finland
first.last@helsinki.fi

## Abstract

We present our work on the problem of detection Multiple Admissibility (MA) in language learning. Multiple Admissibility occurs when more than one grammatical form of a word fits syntactically and semantically in a given context. In second-language education—in particular, in intelligent tutoring systems/computer-aided language learning (ITS/CALL), systems generate exercises automatically. MA implies that multiple alternative answers are possible. We treat the problem as a grammaticality judgement task. We train a neural network with an objective to label sentences as grammatical or ungrammatical, using a "simulated learner corpus": a dataset with correct text and with artificial errors, generated automatically. While MA occurs commonly in many languages, this paper focuses on learning Russian. We present a detailed classification of the types of constructions in Russian, in which MA is possible, and evaluate the model using a test set built from answers provided by users of the Revita language learning system.

## 1 Introduction

The problem of *Multiple Admissibility* (MA) occurs in the context of language learning. In "cloze" exercises (fill-in-the-blank), the learner receives a text with some word removed, and a base form[1] of the removed word as a hint. The task is to produce the correct grammatical form of the missing word, given the context. The answer given by the user is checked automatically by the language learning system. Therefore, the system should be able to accept more than one answer, if there are grammatically and semantically valid alternatives in the given context. Otherwise, the language learning system returns negative (actually incorrect) feedback to the learner. This is a problem, because negative feedback for an acceptable answer misleads and discourages the learner.

We examine MA in the context of our language learning system, Revita (Katinskaia et al., 2018). Revita is available online[2] for second language (L2) learning beyond the beginner level. It is in use in official university-level curricula at several major universities. It covers several languages, many of which are highly inflectional, with rich morphology. Revita creates a variety of exercises based on input text materials, which are selected by the users. It generates exercises and assesses the users' answers automatically.

For example, consider the sentence in Finnish: "Ilmoitus *vaaleista* *tulee* *kotiin* *postissa*." ("Notice about elections comes to the house in the mail.")

In practice mode, Revita presents the text to the learner in small pieces—"snippets" (about 1 paragraph each)—with all generated exercises. This is important, because grammatical forms in exercises usually depend on a wider context.

For the given example, Revita can generate cloze exercises hiding several tokens (*surface forms*) and providing their lemmas as hints:

"Ilmoitus *vaali* *tulla* *koti* *posti*."

("Notice election come house mail.").

For the verb "*tulla*" ("come") and nouns "*vaali*" ("election"), "*koti*" ("home") and "*posti*" ("mail") the learner should insert the correct grammatical forms. Revita expects them to be the same as the forms in the original text, and will return negative feedback otherwise. However, in this example, *postitse* ("via email") is also acceptable, although it is not in the original text.

The MA problem is also relevant in the context of exercises that require a free-form answer, such as an answer to a question or an essay. The learner

[1] The base or "dictionary" form will be referred to as *lemma* in this paper.

[2] https://revita.cs.helsinki.fi/



*Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 12–22,
Florence, Italy, 2 August 2019. 

can produce "unexpected" but nevertheless valid grammatical forms.

In the current work, we restrict our focus to MA of multiple surface forms of the same lemma, given the context; we do not consider synonyms, which can also fit the same context. The latter topic is part of the future work.

The structure of the paper is as follows: In section 2 we formulate the problem and provide a detailed classification of the types of MA found in the learner data. In section 3 we describe our approach, in particular, the procedure for generating artificial grammatical errors and creating test sets. Section 4 presents previous work on artificial error generation. Section 5 describes our model and experimental setup. In section 6 we discuss the results and error analysis, and conclude in section 7.

## 2 Multiple Admissibility: Problem Overview

We use data from several hundred registered students learning Russian (and other languages), practicing with exercises based on texts, or answering questions in test sessions. Currently, Revita does not provide a mode for written essays, because it does not check free-form answers.

Except for cloze exercises, Revita also generates exercises where the users are asked to select among various surface forms based on a given lemma—a correct surface form and a set of automatically generated distractors; or to type the word they hear. This allows us to collect a wide range of errors, though not all kinds of possible errors; e.g., currently we do not track punctuation errors, word order errors, insertion errors, errors resulting from choosing an incorrect lemma, etc.

We annotated 2884 answers from the Revita database, which were automatically marked as "not matching the original text." This work was done by two annotators (90% agreement) both with native-level competency in Russian, and background in linguistics and teaching Russian. Among all annotated answers, 7.5% were actually correct, but differed from the original text. Also, we checked answers given by three students with C1 CEFR proficiency level in Russian (established independently by their teachers); 15.8% of these answers were grammatically and semantically valid in the context. Thus, for advanced users, the problem of MA is twice as relevant as on average; we plan to investigate these results with a larger base of Revita users.

### 2.1 Types of Multiple Admissibility

In analyzing the answers given by our users, we discovered several types of the most frequent contexts where MA appears.

**Present/Past tense**: The most clear case of MA in Russian (as in many other languages) is the case of interchangeable forms of present and past tense of verbs. Russian has three tenses (present, past, future), of which past and future tenses can have perfective or imperfective aspect.

In the next example, if both verbs are chosen for as exercises[3] and the learner sees two lemmas ("задержаться" and "вернуться"), she has a possibility to produce verb forms in any tense depending on the context *beyond the given sentence*. It may be narrative past tense or present tense, or the text may be a message where the communicative goal of the speaker is to inform the reader about future events, so future tense is expected.

> "Мы задержались(PST)[4] у друзей и вернулись(PST) домой поздно."
> "We were delayed with friends and returned home late."

The following option may be acceptable:

> "Мы задержались(PST) у друзей и вернемся(FUT) домой поздно."
> "We were delayed with friends and will return home late."

The future cannot precede the past in this sentence, so the next variant answer is grammatically incorrect:

> * "Мы задержимся(FUT) у друзей и вернулись(PST) домой поздно."
> "We will be delayed with friends and returned home late."[5]

Some cases are more difficult because the choice of tense can depend on knowledge beyond the text:

> "Ученым удалось установить, что у минойцев существовало (PST) несколько видов письма."

[3] If one of the verbs is chosen as an exercise, the user may get a hint from the surface form of the other verb, that they should be coordinated.

[4] We use standard abbreviations from the Leipzig Glossing Rules.

[5] *Star is used to mark an incorrect sentence.

"Scientists were able to establish that the Minoans had several types of writing systems."

In a non-fictional narrative, only the past tense can be used, since the Minoans do not exist in the present. These examples show that deciding which tenses are acceptable in the context is difficult.

**Singular/Plural**: Singular and plural nouns can be grammatically valid in the same context, if there is no dependency on words beyond the sentence boundaries. For instance:

"Из последних разработок—скафандр для работы (SG.) в открытом космосе."

"Из последних разработок—скафандр для работ (Pl.) в открытом космосе."

"From the latest developments—the spacesuit for work in open space."

**Short/Full adjective**:[6] in many constructions, short and full forms of adjectives can be used as a part of a compound predicate (Vinogradov, 1972). The difference between these is that the short form typically expresses *temporal* meaning and that it is a phenomenon of *literary language*, whereas its full alternative form sounds more colloquial.

"Вы ужасно болтливы (short) и непоседливы " (short) (from I. Bunin)

"You are being terribly talkative (PRED) and restless (PRED)."

"Вы ужасно болтливые и непоседливые".

"You are terribly talkative (DESCR) and restless (DESCR)".

We treat these examples as MA, because even for native Russian speakers in many cases the choice can be unclear. So it would be too strict to treat one of the variants as incorrect.

**Nominative/Instrumental case**: nouns as part of compound named predicates can be in the nominative or instrumental case. The difference in meaning is similar to that of short/full adjectives (Rosenthal et al., 1994): nominative indicates a *constant* feature of a subject, whereas instrumental indicates a *temporary* feature of a subject. We consider the following examples as MA, because of a subtle difference in meaning:

"Она была загадочная (NOM) и непонятная (NOM) для меня". (from I. Bunin)

"Она была загадочной (INS) и непонятной (INS) для меня".

"She was mysterious and incomprehensible to me".

**Genitive/Accusative case**: usage of genitive vs. accusative is a complex topic, beyond the scope of this paper. We mention a few examples briefly, where MA can appear—denotation of *a part of the whole* and *negation*. Usually the genitive is used to denote a part of the whole. In the following example, usage of the accusative is incorrect:

"Пожалуйста, отрежь хлеба" (GEN)

* "Пожалуйста, отрежь хлеб" (ACC)

"Please, cut some bread."

In some contexts both meanings are possible—of a part and of the whole—resulting in MA:

"Нам оставили хлеба и вина" (GEN)

"We were left with some bread and wine."

"Нам оставили хлеб и вино" (ACC)

"We were left with bread and wine."

If both words appear in exercises, Revita should accept genitive or accusative (if the context specifies the expected meaning nowhere else).

The next case is negation constructions. The genitive is usually used where negation is stressed, whereas the accusative weakens the negation, (Rosenthal et al., 1994). However, it is worth noticing that the difference can be difficult to understand even for native Russian speakers.

"До вас никто еще этого браслета (GEN) не надевал." (from Kuprin.)

"**No one** has worn this bracelet before you."

Compare with a similar example in the accusative:

"Он не отвергнул тогда с презрением эти сто (ACC) рублей." (Dostoevsky)

"He **did not** reject those one hundred rubles with contempt."

It is not always possible for the learner to know which case expected in the sentence, because it implies that she should know which type of negation was mentioned. always possible, and both options fit semantically.

**Perfective/Imperfective aspect**: errors in aspect are very common (Rozovskaya and Roth,

[6] So-called full/short forms of adjectives correspond to descriptive/predicative adjectives in other languages: compare "the hungry(DESCR) dog" vs. "the dog is hungry(PRED)".

2019). Without going into detail, we show examples from (Rakhilina et al., 2014)—a heritage learner corpus.[7] Both sentences can be interpreted as correct, with a subtle difference in meaning:

> "Он пишет ей, чтобы она не перестала (PFV) любить его."
>
> "He writes to her that she should not stop loving him."
>
> "Он пишет ей, чтобы она не переставала (IPFV) любить его."
>
> "He writes to her that she continue to love him."

**Gerund/Other verb forms**: MA occurs in some contexts where gerunds are used. In the following example, both sentences can express the meaning of two actions happening at the same time. The only argument against using a past tense verb form is that it sounds somewhat unnatural without a conjunction "and" between verbs ("was saying and thinking"):

> "... говорил я себе, думая (GERUND) об Охотном ряде" (from Bunin)
>
> "... I was saying to myself, thinking about Okhotny Ryad"
>
> "... говорил я себе, думал (PST) об Охотном ряде"
>
> "... I was saying to myself, was thinking about Okhotny Ryad"

**Prepositions with multiple cases**: some prepositions can govern two different cases of the following noun, with no change in meaning:

> "Она спряталась под одеялом." (INS)
>
> "Она спряталась под одеяло." (ACC)
>
> "She hid under a blanket."

**Second Genitive (Partitive)**: and other forms common in spoken language can be valid alternatives, often unfamiliar to L2 learners.

> "Я привозил ей коробки шоколаду (2GEN), новые книги..."
>
> "Я привозил ей коробки шоколада (GEN), новые книги..."
>
> "I was bringing her boxes of chocolate, new books..."

**Other cases**: some examples of MA contexts are exceptionally interesting and rare.

> "На этой почве хорошо росла трава, что обеспечивало (Neu) пастбищами овец."
>
> "На этой почве хорошо росла трава, что обеспечивала (Fem) пастбищами овец."
>
> "The grass grew well on this soil, what provided sheep with pasture."

These sentences expresses similar meaning, although the verb "обеспечивать" ("*to provide*") appears in the neuter gender in the first and feminine in the second. This happens because in the first sentence the subordinative pronoun "что" ("*which*") refers to the entire preceding clause, whereas in the second it refers only to the word "трава" ("*grass*"), which is feminine.

We observe many other types of constructions with MA. This is not an exhaustive list, but covers only some of the types actually found among the answers given by the learners of Russian in Revita. The list should give us some intuitions about the problem of MA, and how difficult it is to identify automatically.

## 3 Overview of the Approach

How can we identify instances of Multiple Admissibility? One approach to this problem is to train a model with a language modeling objective—referred to as "LM-trained" in the literature. In such a scenario, the task of the model is to predict the next word at every point in the sentence, e.g., for the sentence "The keys to the cabinet [is/are] here"[8] the task of the model is to predict that $P(are|C) > P(is|C)$, where $C$ is the context. Linzen (2016) experimented with this kind of language modeling in three setups: without any grammatically relevant supervision, with supervision on grammaticality (predicting which of two sentences[9] is grammatical/ungrammatical), and number prediction—predicting between two classes, "singular" or "plural". The last two setups are strongly supervised. The poorest results were obtained using a LM-trained model, despite using a large-scale language model (Jozefowicz et al., 2016).

Later, Gulordava (2018) reevaluated these results for the task of predicting long-distance agreement: for several languages, including Russian,

[7] Heritage learners are persons with a cultural connection to or proficiency in the language through family, community, or country of origin. (Definition from http://www.cal.org/heritage/research/)

[8] The example is from (Linzen et al., 2016).

[9] "The keys to the cabinet are here" vs. "The keys to the cabinet is here"

an LM-trained RNN approached the accuracy of the supervised models described in (Linzen et al., 2016). Marvin (2018) performed a targeted evaluation of several LMs—N-gram, LM-trained RNN, and Multitask RNN[10]—on the task of detecting errors in several grammatical constructions for English.[11] The results of even the strongest LM varied considerably depending on the syntactic complexity of the construction: 100% accuracy in the case of simple subject-verb agreement, and 0.52% accuracy for subject-verb agreement across an object relative clause (without "that").

In light of these results from prior research, we decided to approach the problem as a supervised grammaticality judgement task—a two-class classification task, (Linzen et al., 2016)).

Since MA answers are *correct* answers, sentences with alternative grammatical forms of the same lemma would be grammatically correct. One of the problems with this approach is the lack of annotated training data. The Revita database had only 7156 answers labeled "incorrect" for Russian, at the time when these experiments began. Therefore we generated a training dataset by *simulating* grammatical errors. We describe the simulation procedure in the following subsection, and briefly review prior approaches to generating artificial errors for grammatical error detection and correction tasks (GED/GEC). Every instance in the simulated dataset is labeled as correct or incorrect. The network reads the entire sequence in a bidirectional fashion and receives a supervised signal at the end. We describe the model and the experiments in the following sections.

### 3.1 Generating Artificial Errors

First, we describe the process of generating training data: the source of data, preprocessing steps and a brief analysis of what types of errors we obtain in the simulated data. In the following subsection, we proceed to describe the test sets which were build from real users' data containing "natural" errors.

Generating training datasets with artificial errors is a common approach, because obtaining large error-annotated learner corpora is extremely difficult and costly, (Granger, 2003): difficulties relate to collecting data from language learners and very expensive annotation procedures.

Revita at present creates exercises only for words which do not exhibit *lemma ambiguity* (homography).[12] Lemma ambiguity occurs when a surface form has more than one lemma. An example of this type of ambiguity is the token "стекло", which has two morphological analyses:

стечь ("*to flow down*") Verb+Past+Sing+Neut,
стекло ("*glass*") Noun+Sing+Nom/Acc.[13]

In this setting, we do not need to generate errors for surface forms with lemma ambiguity.

Training instances are generated by sliding a window of radius $r$ over the list of input tokens, with the target token in the middle. The target is every $n$-th token ($n$ is the stride). If the target token is unambiguous, is above a frequency threshold,[14] and has a morphological analysis, it is replaced by a *random* grammatical form from the paradigm to which the token belongs.[15]

We use $r = 10$, which results in a window wider than an average sentence in Russian, and we are interested in including wide context in training instances. All generated windows are labeled as negative/ungrammatical. The training dataset consists of a balanced number of grammatical and ungrammatical instances. Part of the generated data was removed from training dataset and used as a *validation set* for training the model.

Of the automatically generated errors that we checked, some appear very natural, while others may be less likely to be made by real students.

As Linzen (2016) notes, some of the generated instances will *not* in fact be ungrammatical. We analysed 500 randomly chosen generated windows; 3% of them happened to be grammatical (in Table 1 we refer to them as Multi-admissible). We provide the interpretation for all labels in Table 1 in the following subsection.

### 3.2 Test Data Analysis

To create test sets, we took 2884 answers from Revita's database, which were automatically marked as "incorrect," and manually annotated them using the labels below:

[10] Language modeling objective, combined with a supervised task of sequence tagging with CCG supertags.

[11] The authors expected that a LM would assign a higher probability to a grammatical sentence than to an ungrammatical one.

[12] Because it currently does not attempt to perform disambiguation, and only one lemma can be shown as the hint.

[13] This is an example not only of lemma ambiguity, but also of word sense and morphological ambiguity.

[14] We count frequencies from the entire corpus used for building the training set, to exclude words appearing once.

[15] We generate paradigms of inflected words using the pymorphy2 morphological analyzer (Korobov, 2015).

| Label | (1) Training set | (2) Real student data | (3) Advanced students |
|---|---|---|---|
| Grammatical error | 83.0% | 21.4% | 39.4% |
| Non-word error | — | 20.8% | 12.2% |
| Multiple-choice | — | 12.0% | 17.0% |
| **Multi-admissible** | **3.0%** | **7.5%** | **15.8%** |
| Pragmatic error | 2% | 1.6% | 2.9% |
| Broken | 12% | 36.7% | 12.7% |
| Total instances annotated | 500 | 2884 | 170 |

Table 1: Data we annotated for verification and testing: (1) subset of the set of errors *automatically generated* for training (randomly sampled and manually annotated), (2) learners' answers (randomly sampled), marked by the System as *incorrect*, (3) subset of learner' incorrect answers—for *advanced* learners only (CEFR level C1/C2). "Broken": discarded instances (technical problems, too many unknown words, numbers, punctuation marks, etc.)

• **Grammatical error**: answer was a valid grammatical form of the word (exists in paradigm), but incorrect in the given context. This group includes only errors made in cloze exercises.
• **Non-word error**: spelling error—the word was rejected by the morphological analyzer.
• **Multiple-choice**: error in a choice of word from a list of options.
• **Multi-admissible**: as mentioned above, we consider these to be *correct* answers.
• **Pragmatic error**: a separate type of error where the given answer can fit grammatically, but is semantically/pragmatically unnatural in the context.

We provide one example of the last kind of error; it requires further investigation:

> "У меня машина сломалась, и мне пришлось звонить в автосервис (ACC)".
> "My car broke down and I had to call (to) the auto repair".

* "У меня машина сломалась, и мне пришлось звонить в автосервисе (LOC)".

* "My car broke down and I had to call (while being) in a car-service station".

Preposition "в" ("in") governs two cases—Nominative and Locative—but the second sentence does not make sense pragmatically. We have begun a more detailed annotation of all learner answers (i.e., the types of grammatical errors). This topic is beyond the scope of this paper.
• **Broken**: discarded instances (technical problems, words not in our training vocabulary, too many numbers, punctuation marks, answers given in languages different from expected, etc).

Table 1 represents the number of all mentioned data types in the real learners' answers (the second column) and in the subset of these real answers which were given only by advanced learners (the third column).

We separate the real, manually annotated data into four test sets (see Table 2).

**A**. The first test set contains only sentences exhibiting MA.

**B**. The second test set is randomly chosen correct sentences from a separate corpus (for a total of 500 instances) which was not used for generating training data.

**C**. The third test set is made to test the ability of our model to distinguish between grammatical and ungrammatical sentences (as it was trained to do)—thus it contains:

**C1**. sentences with grammatical errors made by Revita users;

**C2**. correct sentences from Revita's database.

**D**. The fourth test set contains sentences only with pragmatic errors.

In the next section, we shortly review prior work related to artificial error generation (AEG) for the grammaticality judgement task.

## 4 Related Work

Felice (2016) divides methods of AEG into deterministic vs. probabilistic. The **deterministic approach** consists of methods that generate errors in systematic ways, which do not make use of learner error distributions. Izumi et al. (2003) introduced a system for correction of article errors made by English learners, native in Japanese. The system was trained on artificial data where *a, an, the* or the zero article were replaced with a different option chosen randomly.

Sjöbergh and Knutsson (2005) created an artificial corpus consisting of two of the most frequent types of errors among non-native Swedish speak-

ers: split compounds and word order errors.

Brockett et al. (2006) describe a statistical machine translation (SMT) system for correcting a set of 14 countable and uncountable nouns which are often confused by learners of English. They used rules to change quantifiers (e.g. *much–many*), to generate plural forms, and to insert unnecessary determiners. Lee and Seneff (2008) created an artificial corpus of verb form errors. They changed verbs in the original text to different forms, such as to-infinitive, 3rd person singular present, past, or -ing participle. Ehsan and Faili (2013) used SMT for AEG to correct grammatical errors and context-sensitive spelling mistakes in English and Farsi. Training corpora were obtained by injecting artificial errors into well-formed treebank sentences using predefined error templates.

**Probabilistic approach:** Rozovskaya and Roth describe several methods for AEG which include creation of article (Rozovskaya and Roth, 2010b) and preposition errors (Rozovskaya and Roth, 2010a, 2011) based on statistics from an English as a Second Language (ESL) corpora. They inject errors into Wikipedia sentences using different strategies (e.g., distribution before and after correction, L1-specific error distributions).

Rozovskaya et al. (2012) proposed an inflation method, which preserves the ability of the model to take into account learner error patterns. While also increasing the model's recall, this method reduced the confidence that the system has in the source word. Improvement in F-scores was achieved by this method when correcting determiners and prepositions. Further, this method was used by other researchers (Felice and Yuan, 2014; Putra and Szabó, 2013; Rozovskaya et al., 2013, 2014, 2017).

Dickinson (2010) introduce an approach to generate artificial syntactic errors and morphological errors for Russian. Imamura et al. (2012) adapt the method of Rozovskaya and Roth (2010b) for particle correction in Japanese. Cahill et al. (2013) examine automatically-compiled sentences from Wikipedia revisions for correcting errors in prepositions. Kasewa et al. (2018) use an off-the-shelf attentive sequence-to-sequence NN (Bahdanau et al., 2014) to learn to introduce errors.

## 5 Model and Experiment

**Data:** For generating the training/validation datasets, we use the open-source "Taiga" Russian corpus,[16] which is arranged by genre into several segments. We used all news segments, and part of the literary text segment, for a total of 809M words. We exclude social media, film subtitles, and poems, because their language has more deviations from the literary standard. All documents were lowercased, tokenized, and morphologically analyzed using Crosslator (Klyshinsky et al., 2011).[17] We replace all punctuation marks with a special token, to preserve information about sentence/clause boundaries. The size of the training vocabulary was around 1.2M words (after removing words with frequency less than 2). For validation, we randomly chose 5% of all generated data.

**Model architecture**: our baseline neural network (NN) is implemented in TensorFlow. Its architecture is a one-layer bidirectional LSTM with dropout (0.2), which has 512 hidden units. The hidden state of the BiLSTM is then fed to an Multi-layer Perceptron (MLP). The MLP uses one hidden layer with 1024 neurons, and Leaky ReLU activation function. The size of the output layer is 1, since we have only two classes to predict. The output of the MLP is then fed to a sigmoid activation function to obtain a prediction for the entire input sequence. To encode words, we use the FastText 300-dimensional pre-trained embeddings.[18]

The network and the word embeddings were trained in an end-to-end fashion. Optimization was done using Adam, dropout, and early stopping based on the loss on the validation set. We trained the network over only half of an epoch, since it was showing signs of overfitting—because we use a sliding window, the number of training instances was over 90M. The averaged accuracy on the validation set was 95 %. Table 2 reports the accuracy on the test sets, averaged across 5 runs.

## 6 Results

Table 2 shows the results of our experiments in terms of accuracy. 85.9% accuracy was achieved across all types of MA. However, we should stress that in the test set marked Multi-admissible (MA), the majority of the instances belong to the MA types of Present/Past tense and Singular/Plural.

Since the test set has a small number of instances of MA contexts with gerund/other verb

[16] https://tatianashavrina.github.io

[17] This analyzer was chosen because it is a part of Revita's text processing pipeline.

[18] https://fasttext.cc/docs/en/crawl-vectors.html

| | *Test set* | *#* | *Acc* |
|---|---|---|---|
| A. | **Multi-admissible** | 178 | 85.9 |
| B. | Random correct | 500 | 92.3 |
| C. | Correct & incorrect | 1290 | 81.0 |
| C1. | Grammatically correct | 650 | 73.3 |
| C2. | Grammatically incorrect | 640 | 88.6 |
| D. | Pragmatic errors | 46 | 54.3 |

Table 2: *Percent accuracy* of our NN model. Random correct: test set built from sentences which were not included in the training and validation sets and did not appear in Revita's database, randomly selected sentences from normal texts. Grammatically incorrect: test set with real grammatical errors from students' data. Pragmatic errors: test set with real pragmatic errors from students' data.

| | *MA types* | *#* | *Acc* |
|---|---|---|---|
| 1. | Perf/Imperf + Gerund/Other | 14 | 92.8 |
| 2. | Case | 24 | 91.7 |
| 3. | Present/Past | 53 | 88.7 |
| 4. | Singular/Plural | 78 | 82.0 |
| 5. | Short/Full adj | 9 | 77.7 |

Table 3: *Percent accuracy* of our NN model for different MA contexts. Case combines all types of MA contexts listed in the Subsection 2.1 which differ by case (Nominative/Instrumental, Genitive/Accusative and others).

forms and MA contexts which differ by perfective/imperfective aspect, we grouped them together for testing the model. On these two types of MA the model achieved the highest accuracy, 92.8% (see Table 3). For the same reasons we grouped together MA contexts which differ by case (Nominative/Instrumental, Genitive/Accusative, Second Genitive and other). The overall accuracy for these contexts is 91.7%.

We plan to test all combined MA types separately as soon as we have more annotated data. The accuracy for Present/Past tense is 88.7%. The accuracy for Number agreement (including subject-verb agreement on number) is 81.0%. The lowest accuracy was achieved for Short/Full adjectives—77.7%. Some discussion of errors is in the following subsection.

We use additional test sets to assess other aspects of the trained NN model. The "Random correct" test set (B.) contains 500 randomly sampled sentences *without errors*, to compare with the MA test set. These sentences were sampled from a corpus which was not used for generating training/validation data, and are not present in the Revita database. On random correct sentences, the model achieved substantially better results than for MA instances (92.3%). It is interesting that the model has more difficulty with the syntactic structure of contexts with known MA than with some random correct contexts.

Another test set (C.) is made up of correct sentences from the Revita database, and sentences with grammatical errors made by the learners. We evaluate the model on this test set to gain insight into how well it can differentiate between various incorrect vs. correct sentences. The discussion of results for different grammatical error types is beyond the scope of this paper, and is left for the further work.

The pragmatic error test set (D.) was used to find out how difficult it is to predict labels for sentences which are correct grammatically but incorrect semantically/pragmatically. Clearly these instances pose the greatest challenge to the current model; (it was not explicitly trained to detect them).

Of the nearly 3000 manually annotated instances, the number of instances found to be pragmatic errors and MA was not large.

### 6.1 Error Analysis

We analysed some of the errors the model made on the MA test dataset. For all types of MA, we found some similar patterns: the model assigns very low scores to short sequences (which are padded), contexts with too many punctuation marks or names, and context with non-Russian words which are unknown to the model. For example, for constructions with Present/Past tense, the network made wrong predictions if the subject was a name or a number, which in most cases corresponds to the token "UNK" in our model's vocabulary. The same happens if the subject is outside the window. Sometimes the model confuses certain nouns or pronouns next to the verb with its subject, for example:

> "Поезда (SUBJ) метро задерживаются (PRED).
> "Trains of the metro are delayed."

In this case, the model might suppose that the (genitive) singular noun "метро" ("metro") is the subject of the plural verb "задерживаются" ("are delayed"), which is incorrect—the actual

subject is the plural noun "поезда" ("trains")—but the genitive is closer to the predicate. As a result, the model will identify the sentence as grammatically incorrect, believing that the subject and predicate conflict in number.

We also should note that some instances marked by the model as incorrect are actually incorrect, but marked as MA by annotators, which means that MA instances need to be double-checked and that the model is able to identify ungrammatical contexts.

It is difficult to compare our results directly with prior work, because we have not yet found in the previous work a problem similar to Multiple Admissibility for Russian. A similar problem—grammatical acceptability judgment—is presented in (Warstadt et al., 2018), for English only. The best results they achieved in terms of percent accuracy is 77.2%. The average human accuracy is 85%.

For the task of grammatical error detection, the results obtained for Russian are much lower than for English. For example, the highest precision in (Rozovskaya and Roth, 2019) for errors in number agreement is 56.7.

Concerning the grammaticality judgement task, Marvin (2018) reported accuracies for subject-verb agreement from 50% to 99% depending on the syntactic complexity of the sentence (e.g., relations across relative clause). This is similar to Present/Past tense construction in our setup.

Linzen et al. (2016) also concludes that the grammaticality judgment objective is more difficult than, for example, the number prediction objective. The LSTM model can have up to 23% error on this task, as sentence complexity grows. This work studied only number agreement and subject-verb dependencies in English.

## 7 Conclusions and Future Work

We address the problem of *Multiple admissibility* in automatically generated exercises, by approaching it as a grammaticality judgment task. We offer a detailed study of examples, where our language learning system mistakenly assesses admissible answers as incorrect. We classify these contexts into 10 types, where only some of these types have been in the focus of prior research, especially for Russian. We train a NN model with the grammaticality objective, independent of the type of test set we use for evaluation. The problem of lacking labeled training data was approached by generating a dataset with artificial errors. We also observed that the MA problem is more relevant for advanced language learners. Another observation is that for a trained model it is more difficult to make prediction about MA contexts than about random correct sentences.

We plan to extend and improve our training data by marking numbers with special tokens, or by mapping them into words. We also plan to mark names with a name tag by using some of the existing NER models, and mark rare words with their part of speech. We also believe that providing the model with syntactic information (parsing) can help, so that we can train a model in a multitask fashion: predict tags of words, as well as their correctness. Also, it is worth trying to use new large-scale language models, which proved to be more effective on a variety of tasks.

Additional annotation of student data collected in Revita's database is needed; in the current work, the annotation was done by two experts, and all disagreements were resolved. We plan to extend our experiments to other languages available in Revita. Each has its own language-specific types of MA. Generating all paradigms of a word could be problematic for some highly inflected languages (e.g., Finnish, etc.).

The goal of this paper is to introduce the problem of Multiple Admissibility and to attract more attention to experimenting with morphologically rich languages and languages other than English in this context.

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Chris Brockett, William B Dolan, and Michael Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 249–256. Association for Computational Linguistics.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust systems for preposition error correction using Wikipedia revisions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 507–517.

Markus Dickinson. 2010. Generating learner-like morphological errors in russian. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 259–267.

Nava Ehsan and Heshaam Faili. 2013. Grammatical and context-sensitive error correction using a statistical machine translation framework. *Software: Practice and Experience*, 43(2):187–206.

Mariano Felice. 2016. Artificial error generation for translation-based grammatical error correction. Technical report, University of Cambridge, Computer Laboratory.

Mariano Felice and Zheng Yuan. 2014. Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–126.

Sylviane Granger. 2003. Error-tagged learner corpora and CALL: A promising synergy. *CALICO journal*, 20(3):465–480.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. 2012. Grammar error correction using pseudo-error sentences and domain adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 388–392. Association for Computational Linguistics.

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. 2003. Automatic error detection in the Japanese learners' English spoken data. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. 2018. Wronging a right: Generating better errors to improve grammatical error detection. *arXiv preprint arXiv:1810.00668*.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of LREC: 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

E.S. Klyshinsky, N.A. Kochetkova, M.I. Litvinov, and V.Yu. Maximov. 2011. Method of POS-disambiguation using information about words co-occurrence (for Russian). *Proceedings of GSCL*, pages 191–195.

Mikhail Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 320–332. Springer.

John Lee and Stephanie Seneff. 2008. Correcting misuse of verb forms. *Proceedings of ACL-08: HLT*, pages 174–182.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Desmond Darma Putra and Lili Szabó. 2013. UdS at CoNLL 2013 Shared Task. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 88–95.

E.V. Rakhilina, A.S. Vyrenkova, and M.S. Polinskaya. 2014. Grammar of errors and grammar of constructions: "Heritage" ("inherited") Russian language. *Questions of linguistics*, 3(2014):3–19.

Ditmar Elyashevich Rosenthal, E.V. Dzhandzhakova, and N.P. Kabanova. 1994. *Reference on Spelling, Pronunciation, and Literary Editing*. Moscow International School of Translators.

Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, and Dan Roth. 2013. The University of Illinois system in the CoNLL-2013 shared task. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 13–19.

Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, Dan Roth, and Nizar Habash. 2014. The Illinois-Columbia system in the CoNLL-2014 shared task. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 34–42.

Alla Rozovskaya and Dan Roth. 2010a. Generating confusion sets for context-sensitive error correction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 961–970. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2010b. Training paradigms for correcting errors in grammar and usage. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pages 154–162. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of*

*the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 924–933. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

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

Alla Rozovskaya, Mark Sammons, and Dan Roth. 2012. The UI system in the HOO 2012 shared task on error correction. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 272–280. Association for Computational Linguistics.

Jonas Sjöbergh and Ola Knutsson. 2005. Faking errors to avoid making errors: Very weakly supervised learning for error detection in writing. In *Proc. RANLP*, volume 2005.

Viktor Vladimirovich Vinogradov. 1972. *Russian language: The grammatical doctrine of the word*, volume 21. High School.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.