Czech Grammar Error Correction with a Large and Diverse Corpus

Jakub Náplava[†] **Milan Straka**[†] **Jana Straková**[†] **Alexandr Rosen**[‡]

[†]Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{naplava,straka,strakova}@ufal.mff.cuni.cz

[‡]Charles University, Faculty of Arts
Institute of Theoretical and Computational Linguistics
alexandr.rosen@ff.cuni.cz

Abstract

We introduce a large and diverse Czech corpus annotated for grammatical error correction (GEC) with the aim to contribute to the still scarce data resources in this domain for languages other than English. The *Grammar Error Correction Corpus for Czech* (*GECCC*) offers a variety of four domains, covering error distributions ranging from high error density essays written by non-native speakers, to website texts, where errors are expected to be much less common. We compare several Czech GEC systems, including several Transformer-based ones, setting a strong baseline to future research. Finally, we meta-evaluate common GEC metrics against human judgements on our data. We make the new Czech GEC corpus publicly available under the CC BY-SA 4.0 license at http://hdl.handle.net/11234/1-4639.

1 Introduction

Representative data both in terms of size and domain coverage are vital for NLP systems development. However, in the field of grammar error correction (GEC), most GEC corpora are limited to corrections of mistakes made by foreign or second language learners even in the case of English (Tajiri et al., 2012; Dahlmeier et al., 2013; Yannakoudakis et al., 2011, 2018; Ng et al., 2014; Napoles et al., 2017). At the same time, as recently pointed out by Flachs et al. (2020), learner corpora are only a part of the full spectrum of GEC applications. To alleviate the skewed perspective, the authors released a corpus of website texts.

Despite recent efforts aimed to mitigate the notorious shortage of national GEC-annotated corpora (Boyd, 2018; Rozovskaya and Roth, 2019; Davidson et al., 2020; Syvokon and Nahorna, 2021; Cotet et al., 2020; Náplava and Straka, 2019), the lack of adequate data is even more acute

in languages other than English. We aim to address both the issue of scarcity of non-English data and the ubiquitous need for broad domain coverage by presenting a new, large and diverse Czech corpus, expertly annotated for GEC.

Grammar Error Correction Corpus for Czech (*GECCC*) includes texts from multiple domains in a total of 83 058 sentences, being, to our knowledge, the largest non-English GEC corpus, as well as being one of the largest GEC corpora overall.

In order to represent a diversity of writing styles and origins, besides essays of both native and non-native speakers from Czech learner corpora, we also scraped website texts to complement the learner domain with supposedly lower error density texts, encompassing a representation of the following four domains:

- *Natives Formal* – essays written by native students of elementary and secondary schools
- *Natives Web Informal* – informal website discussions
- *Romani* – essays written by children and teenagers of the Romani ethnic minority
- *Second Learners* – essays written by non-native learners

Using the presented data, we compare several state-of-the-art Czech GEC systems, including some Transformer-based.

Finally, we conduct a meta-evaluation of GEC metrics against human judgements to select the most appropriate metric for evaluating corrections on the new dataset. The analysis is performed across domains, in line with Napoles et al. (2019).

Our contributions include (i) a **large and diverse Czech GEC corpus**, covering learner corpora and website texts, with unified and, in some domains, completely new GEC annotations, (ii) a **comparison of Czech GEC systems**, and (iii) a **meta-evaluation of common GEC metrics** against human judgement on the released corpus.

2 Related Work

2.1 Grammar Error Correction Corpora

Until recently, attention has been focused mostly on English, while GEC data resources for other languages were in short supply. Here we list a few examples of English GEC corpora, collected mostly within an English-as-a-second-language (ESL) paradigm. For a comparison of their relevant statistics see Table 1.

Lang-8 Corpus of Learner English (Tajiri et al., 2012) is a corpus of English language learner texts from the Lang-8 social networking system.

NUCLE (Dahlmeier et al., 2013) consists of essays written by undergraduate students of the National University of Singapore.

FCE (Yannakoudakis et al., 2011) includes short essays written by non-native learners for the Cambridge ESOL First Certificate in English.

W&I+LOCNESS is a union of two datasets, the *W&I (Write & Improve)* dataset (Yannakoudakis et al., 2018) of non-native learners' essays, complemented by the *LOCNESS* corpus (Granger, 1998), a collection of essays written by native English students.

The GEC error annotations for the learner corpora above were distributed with the BEA-2019 Shared Task on Grammatical Error Correction (Bryant et al., 2019).

The *CoNLL-2014* shared task test set (Ng et al., 2014) is often used for GEC systems evaluation. This small corpus consists of 50 essays written by 25 South-East Asian undergraduates.

JFLEG (Napoles et al., 2017) is another frequently used GEC corpus with fluency edits in addition to usual grammatical edits.

To broaden the restricted variety of domains, focused primarily on learner essays, a *CWEB* collection (Flachs et al., 2020) of website texts was recently released, aiming at contributing lower error density data.

AESW (Daudaravicius et al., 2016) is a large corpus of scientific writing (over 1M sentences), edited by professional editors.

Finally, Napoles et al. (2019) recently released *GMEG*, a corpus for the evaluation of GEC metrics across domains.

Grammatical error correction corpora for languages other than English are less common and – if available – usually limited in size and domain: German *Falko-MERLIN* (Boyd, 2018), Russian *RULEC-GEC* (Rozovskaya and Roth,

2019), Spanish *COWS-L2H* (Davidson et al., 2020), Ukrainian *UA-GEC* (Syvokon and Nahorna, 2021) and Romanian *RONACC* (Cotet et al., 2020).

To better account for multiple correction options, datasets often contain several reference sentences for each original noisy sentence in the test set, proposed by multiple annotators. As we can see in Table 1, the number of annotations typically ranges between 1 and 5 with an exception of the CoNLL14 test set, which – on top of the official 2 reference corrections – later received 10 annotations from Bryant and Ng (2015) and 8 alternative annotations from Sakaguchi et al. (2016).

2.2 Czech Learner Corpora

By the early 2010s, Czech was one of a few languages other than English to boast a series of learner corpora, compiled under the umbrella project *AKCES*, evoking the concept of *acquisition corpora* (Šebesta, 2010).

The native section includes transcripts of handwritten essays (*SKRIPT 2012*) and classroom conversation (*SCHOLA 2010*) from elementary and secondary schools. Both have their counterparts documenting the Roma ethnolect of Czech:[1] essays (*ROMi 2013*) and recordings and transcripts of dialogues (*ROMi 1.0*).[2]

The non-native section goes by the name of *CzeSL*, the acronym of *Czech as the Second Language*. *CzeSL* consists of transcripts of short handwritten essays collected from non-native learners with various levels of proficiency and native languages, mostly students attending Czech language courses before or during their studies at a Czech

---

[1] The Romani ethnolect of Czech is the result of contact with Romani as the linguistic substrate. To a lesser (and weakening) extent the ethnolect shows some influence of Slovak or even Hungarian, because most of its speakers have roots in Slovakia. The ethnolect can exhibit various specifics across all linguistic levels. However, nearly all of them are complementary with their colloquial or standard Czech counterparts. A short written text, devoid of phonological properties, may be hard to distinguish from texts written by learners without the Romani backround. The only striking exception are misspellings in contexts where the latter benefit from more exposure to written Czech. The typical example is the omission of word boundaries within phonological words, e.g. between a clitic and its host. In other respects, the pattern of error distribution in texts produced by ethnolect speakers is closer to native rather than foreign learners (Bořkovcová, 2007, 2017).

[2] A more recent release *SKRIPT 2015* includes a balanced mix of essays from *SKRIPT 2012* and *ROMi 2013*. For more details and links see http://utkl.ff.cuni.cz/akces/.

Language	Corpus	Sentences	Err. r.	Domain	# Refs.
English	*Lang-8*	1 147 451	14.1%	SL	1
	NUCLE	57 151	6.6%	SL	1
	FCE	33 236	11.5%	SL	1
	W&I+LOCNESS	43 169	11.8%	SL, native students	5
	CoNLL-2014 test	1 312	8.2%	SL	2,10,8
	JFLEG	1 511	—	SL	4
	GMEG	6 000	—	web, formal articles, SL	4
	AESW	over 1M	—	scientific writing	1
	CWEB	13 574	~2%	web	2
Czech	*AKCES-GEC*	47 371	21.4%	SL essays, Romani ethnolect of Czech	2
German	*Falko-MERLIN*	24 077	16.8%	SL essays	1
Russian	*RULEC-GEC*	12 480	6.4%	SL, heritage speakers	1
Spanish	*COWS-L2H*	12 336	—	SL, heritage speakers	2
Ukrainian	*UA-GEC*	20 715	7.1%	natives/SL, translations and personal texts	2
Romanian	*RONACC*	10 119	—	native speakers transcriptions	1

Table 1: Comparison of GEC corpora in size, token error rate, domain and number of reference annotations in the test portion. SL = second language learners.

university. There are several releases of *CzeSL*, which differ mainly to what extent and how the texts are annotated (Rosen et al., 2020).[3]

More recently, hand-written essays have been transcribed and annotated in *TEITOK* (Janssen, 2016),[4] a tool combining a number of corpus compilation, annotation and exploitation functionalities.

Learner Czech is also represented in *MERLIN*, a multilingual (German, Italian and Czech) corpus built in 2012–2014 from texts submitted as a part of tests for language proficiency levels (Boyd et al., 2014).[5]

Finally, *AKCES-GEC* (Náplava and Straka, 2019) is a GEC corpus for Czech created from the subset of the above mentioned *AKCES* resources (Šebesta, 2010): the *CzeSL-man* corpus (non-native Czech learners with manual annotation) and a part of the *ROMi* corpus (speakers of the Romani ethnolect).

Compared to the *AKCES-GEC*, the new *GECCC* corpus contains much more data (47 371 sentences vs. 83 058 sentences, respectively), by extending data in the existing domains and also adding two new domains: essays written by native learners and website texts, making it the largest

non-English GEC corpus and one of the largest GEC corpora overall.

3 Annotation

3.1 Data Selection

We draw the original uncorrected data from the following Czech learner corpora or Czech websites:

- *Natives Formal* – essays written by native students of elementary and secondary schools from the *SKRIPT 2012* learner corpus, compiled in the *AKCES* project
- *Natives Web Informal* – newly annotated informal website discussions from Czech Facebook Dataset (Habernal et al., 2013a,b) and Czech news site novinky.cz.
- *Romani* – essays written by children and teenagers of the Romani ethnic minority from the *ROMi* corpus of the *AKCES* project and the *ROMi* section of the *AKCES-GEC* corpus
- *Second Learners* – essays written by non-native learners, from the *Foreigners* section of the *AKCES-GEC* corpus, and the *MERLIN* corpus

Since we draw our data from several Czech corpora originally created in different tools with different annotation schemes and instructions, we re-annotated the errors in a unified manner for the entire development and test set and partially also

---

[3]For a list of *CzeSL* corpora with their sizes and annotation details see http://utkl.ff.cuni.cz/learncorp/.
[4]http://www.teitok.org
[5]https://www.merlin-platform.eu

Dataset	Documents	Selected
AKCES-GEC-test	188	188
AKCES-GEC-dev	195	195
MERLIN	441	385
Novinky.cz	—	2 695
Facebook	10 000	3 850
SKRIPT 2012	394	167
ROMi	1 529	218

Table 2: Data resources for the new Czech GEC corpus. The second column (Selected) shows the size of the selected subset from all available documents (first column, Documents).

for the training set.

The data split was carefully designed to maintain representativeness, coverage and backwards compatibility. Specifically, (i) test and development data contain roughly the same amount of annotated data from all domains, (ii) original *AKCES-GEC* dataset splits remain unchanged, (iii) additional available detailed annotations such as user proficiency level in *MERLIN* were leveraged to support the split balance. Overall, the main objective was to achieve a representative cover over development and testing data. Table 2 presents the sizes of data resources in the number of documents. The first column (Documents) shows the number of all available documents collected in an initial scan. The second column (Selected) is a selected subset from the available documents, due to budgetary constraints and to achieve a representative sample over all domains and data portions. The relatively higher number of documents selected for the *Natives Web Informal* domain is due to its substantially shorter texts, yielding fewer sentences; also, we needed to populate this part of the corpus as a completely new domain with no previously annotated data.

To achieve more fine-grained balancing of the splits, we used additional metadata where available: user's proficiency levels and origin language from *MERLIN* and the age group from *AKCES*.

3.2 Preprocessing

De/tokenization is an important part of data preprocessing in grammar error correction. Some formats, such as the M^2 format (Dahlmeier and Ng, 2012), require tokenized formats to track and evaluate correction edits. On the other hand, detokenized text in its natural form is required for other

applications. We therefore release our corpus in two formats: a tokenized M^2 format and detokenized format aligned at sentence, paragraph and document level. As part of our data is drawn from earlier, tokenized GEC corpora *AKCES-GEC* and *MERLIN*, this data had to be detokenized. A slightly modified Moses detokenizer[6] is attached to the corpus. To tokenize the data for the M^2 format, we use the UDPipe tokenizer (Straka et al., 2016).

3.3 Annotation

The test and development sets in all domains were annotated from scratch by five in-house expert annotators,[7] including re-annotations of the development and test data of the earlier GEC corpora to achieve a unified annotation style. All the test sentences were annotated by two annotators; one half of the development sentences received two annotations and the second half one annotation. The annotation process took about 350 hours in total.

The annotation instructions were unified across all domains: The corrected text must not contain any grammatical or spelling errors and should sound fluent. Fluency edits are allowed if the original is incoherent. The entire document was given as a context for the annotation. Annotators were instructed to remove too incomprehensible documents or those containing private information.

To keep the annotation process simple for the annotators, the sentences were annotated (corrected) in a text editor and postprocessed automatically to retrieve and categorize the GEC edits by the ERRor ANnotation Toolkit (ERRANT) (Bryant et al., 2017).

3.4 Train Data

The first source for the training data are the data from the *SKRIPT 2012*; the *MERLIN* corpus and the *AKCES-GEC* train set that were not annotated, thus containing original annotations. These data cover the *Natives Formal*, the *Romani* and the *Second Learners* domain. The second part of the training data are newly annotated data. Specifically, these are all *Natives Web Informal* data and also a small part in the *Second Learners* domain.

---

[6]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/detokenizer.perl

[7]Our annotators are senior undergraduate students of humanities, regularly employed for various annotation efforts at our institute.

All data in the training set were annotated with one annotation.

3.5 Corpus Alignment

The majority of models proposed for grammatical error correction operates over sentences. However, preliminary studies on document-level grammatical error correction recently appeared (Chollampatt et al., 2019; Yuan and Bryant, 2021). The models were shown to benefit from larger context as certain errors such as errors in articles or tense choice do require larger context. To simplify future work with our dataset, we release three alignment levels: (i) sentence-level, (ii) paragraph-level and (iii) document-level. Given that the state-of-the-art grammatical error correction systems still operate on sentence level despite the initial attempts with document-level systems, we perform model training and evaluation at the usual sentence level.[8]

3.6 Inter-Annotator Agreement

As suggested by Rozovskaya and Roth (2010), followed later by Rozovskaya and Roth (2019) and Syvokon and Nahorna (2021), we evaluate inter-annotator agreement by asking a second annotator to judge the need for a correction in a sentence already annotated by someone else, in a single-blind setting as to the status of the sentence (corrected/uncorrected).[9] Five annotators annotated the first pass and three annotators judged the sentence correctness in the second pass. In the second pass, each of the three annotators judged a disjoint set of 120 sentences. Table 3 summarizes the inter-annotator agreement based on second-pass judgements: the numbers represent the percentage of sentences judged correct in the second pass.

Both the average and the standard deviation (82.96 ± 12.12) of our inter-annotator agreement are similar to inter-annotator agreement measured on English (63 ± 18.46, Rozovskaya and Roth 2010), Russian (80 ± 16.26, Rozovskaya and Roth 2019) and Ukrainian (69.5 ± 7.78 Syvokon and Nahorna 2021).

First → Second ↓	A1	A2	A3	A4	A5
A1	—	93.39	97.96	89.63	72.50
A2	84.43	—	95.91	90.18	78.15
A3	68.80	87.68	—	79.39	57.50

Table 3: Inter-annotator agreement based on second-pass judgements: numbers represent percentage of sentences judged correct in second-pass proofreading. Five annotators annotated the first pass, three annotators judged the sentence correctness in the second pass.

3.7 Error Type Analysis

To retrieve and categorize the correction edits from the erroneous-corrected sentence pairs, ERRor ANnotation Toolkit (ERRANT) (Bryant et al., 2017) was used. Inspired by Boyd (2018), we adapted the original English error types to the Czech language. For the resulting set see Table 4. The POS error types are based on the UD POS tags (Nivre et al., 2020) and may contain an optional *:INFL* subtype when the original and the corrected words share a common lemma. The word-order error type was extended by an optional *:SPELL* subtype to allow for capturing word order errors including words with minor spelling errors. The original orthography error type *ORTH* covering both errors in casing and whitespaces is now subtyped with *:WSPACE* and *:CASING* to better distinguish between the two phenomena. Finally, we add two error types specific to Czech: *DIACR* for errors in either missing or redundant diacritics and *QUOTATION* for wrongly used quotation marks. Two original error types remain unchanged: *MORPH*, indicating replacement of a token by another with the same lemma but different POS, and *SPELL*, indicating incorrect spelling.

For part-of-speech tagging and lemmatization we rely on UDPipe (Straka et al., 2016).[10] The word list for detecting spelling errors comes from MorfFlex (Hajič et al., 2020).[11]

We release the Czech ERRANT at `https://github.com/ufal/errant_czech`. We assume that it is applicable to other languages with similar set of errors, especially Slavic languages, if lemmatizer, tagger and morphological dictionary are available.

---

[8]Note that even if human evaluation in Section 5 is performed on sentence-aligned data, human annotators process whole documents, and thus take the full context into account.

[9]A sentence-level agreement on sentence correctness is generally preferred in GEC annotations to an exact inter-annotator match on token edits, since different series of corrections may possibly lead to a correct sentence (Bryant and Ng, 2015).

[10]Using the czech-pdt-ud-2.5-191206.udpipe model.

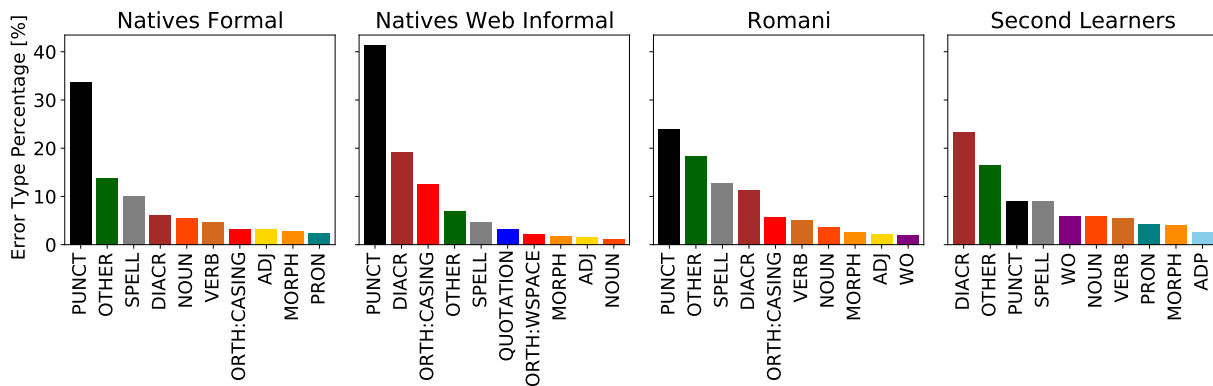[11]We also use the *aggresive* variant of the stemmer from `https://research.variancia.com/czech_stemmer/`.

Figure 1: Distribution of top-10 ERRANT error types per domain in the development set.

Error Type	Subtype	Example
POS (15)		*tažené → řízené*
	:INFL	*manželka → manželkou*
MORPH		*maj → mají*
ORTH	:CASING	*usa → USA*
	:WSPACE	*přes to → přesto*
SPELL		*ochtnat → ochutnat*
WO		*plná jsou → jsou plná*
	:SPELL	*blískají zeleně → zeleně blýskají*
QUOTATION		*" → „*
DIACR		*tiskarna → tiskárna*
OTHER		*sem → jsem ho*

Table 4: Czech ERRANT Error Types.

3.8 Final Dataset

The final corpus consists of 83 058 sentences and is distributed in two formats: the tokenized M^2 format (Dahlmeier and Ng, 2012) and the detokenized format with alignments at the sentence, paragraph and document levels. Although the detokenized format does not include correction edits, it does retain full information about the original spacing.

The statistics of the final dataset are presented in Table 5. The individual domains are balanced on the sentence level in the development and testing sets, each of them containing about 8 000 sentences. The number of paragraphs and documents varies: on average, the *Natives Web Informal* domain contains less than 2 sentences per document, while the *Natives Formal* domain more than 20.

As expected, the domains differ also in the error rate, i.e. the proportion of erroneous tokens (see Table 5). The students' essays in the *Natives Formal* domain are almost 3 times less erroneous than any other domain, while in the *Romani* and *Second Learners* domain, approximately each 4-th token is incorrect.

Furthermore, the prevalence of error types differs for each individual domain. The 10 most common error types in each domain are presented in Figure 1. Overall, errors in punctuation (*PUNCT*) constitute the most common error type. They are the most common error in three domains, although their relative frequency varies. We further estimated that of these errors, 9% (*Natives Formal*) – 27% (*Natives Web Informal*) are uninteresting from the linguistic perspective, as they are only omissions of the sentence formal ending, probably purposeful in case of *Natives Web Informal*. The rest (75–91%) appears in a sentence, most of which (35–68% *Natives Formal*) is a misplaced comma: In Czech, syntactic status of finite clauses strictly determine the use of commas in the sentence. Finally, in 5–7% cases of all punctuation errors, a correction included joining two sentences or splitting a sentence into two sentences. Errors in either missing or wrongly used diacritics (*DIACR*), spelling errors (*SPELL*) and errors in orthography (*ORTH*) are also common, with varying frequency across domains.

Compared to the *AKCES-GEC* corpus, the *Grammar Error Correction Corpus for Czech* contains more than 3 times as many sentences in the development and test sets, more than 50% sentences in the training set and also two new domains.

To the best of our knowledge, the newly introduced *GECCC* dataset is the largest among GEC corpora in languages other than English and it is surpassed in size only by the English *Lang-8* and *AESW* datasets. With the exclusion of these two datasets, the *GECCC* dataset contains more sentences than any other GEC corpus currently known to us.

	Sentence-aligned #sentences			Paragraph-aligned #paragraphs			Doc-aligned #docs			Error Rate
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	
Natives Formal	4 060	1 952	1 684	1 618	859	669	227	87	76	5.81%
Natives Web Informal	6 977	2 465	2 166	3 622	1 294	1 256	3 619	1 291	1 256	15.61%
Romani	24 824	1 254	1 260	9 723	574	561	3 247	173	169	26.21%
Second Learners	30 812	2 807	2 797	8 781	865	756	2 050	167	170	25.16%
Total	66 673	8 478	7 907	23 744	3 592	3 242	9 143	1 718	1 671	18.19%

Table 5: Corpus statistics at three alignment levels: sentence-aligned, paragraph-aligned and doc-aligned. Average Error rate was computed on the concatenation of development and test data in all three alignment levels.

4 Model

In this section, we describe five systems for automatic error correction in Czech and analyze their performance on the new dataset. Four of these systems represent previously published Czech work (Richter et al., 2012; Náplava and Straka, 2019; Náplava et al., 2021) and one is our new implementation. The first system is a pre-neural approach, published and available for Czech (Richter et al., 2012), included for historical reasons as a previously known and available Czech GEC tool; the following four systems represent the current state of the art in GEC: they are all neural network architectures based on Transformers, differing in the training procedure, training data or training objective. A comparison of systems, trained and evaluated on English, Czech, German and Russian, with state of the art is given in Table 6.

4.1 Models

We experiment with the following models:

Korektor (Richter et al., 2012) is a pre-neural statistical spellchecker and (occasional) grammar checker. It uses the noisy channel approach with a candidate model that for each word suggests its variants up to a predefined edit distance. Internally, a Hidden Markov Model (Baum and Petrie, 1966) is built. Its hidden states are the variants of words proposed by the candidate model, and the transition costs are determined from three N-gram language models built over word forms, lemmas and part-of-speech-tags. To find an optimal correction, Viterbi algorithm (Forney, 1973) is used.

Synthetic trained (Náplava and Straka, 2019) is a neural-based Transformer model that is trained to translate the original ungrammatical text to a well formed text. The original Transformer

model (Vaswani et al., 2017) is regularised with an additional source and target word dropout and the training objective is modified to focus on tokens that should change (Grundkiewicz and Junczys-Dowmunt, 2019). As the amount of existing annotated data is small, an unsupervised approach with a spelling dictionary is used to generate a large amount of synthetic training data. The model is trained solely on these synthetic data.

AKCES-GEC (AG) finetuned (Náplava and Straka, 2019) is based on *Synthetic trained*, but finetunes its weights on a mixture of synthetic and authentic data from the *AKCES-GEC* corpus, i.e., on data from the *Romani* and *Second Learners* domains. See Table 6 for comparison with state of the art in English, Czech, German and Russian.

GECCC finetuned uses the same architecture as *Synthetic trained*, but we finetune its weights on a mixture of synthetic and (much larger) authentic data from the newly released GECCC corpus. We use the official code of Náplava and Straka (2019) with the default settings and mix the synthetic and new authentic data in a ratio of 2:1.

Joint GEC+NMT (Náplava et al., 2021) is a Transformer model trained in a multi-task setting. It pursues two objectives: (i) to correct Czech and English texts; (ii) to translate the noised Czech texts into English texts and the noised English texts into Czech texts. The source data come from the *CzEng v2.0* corpus (Kocmi et al., 2020) and were noised using a statistical system Kazi-Text (Náplava et al., 2021) that tries to model several most frequently occurring errors such as diacritics, spelling or word ordering. The statistics of the Czech noise were estimated on the new training set, therefore, the system was indirectly trained also on data from *Natives Formal* and *Natives Web Informal* domains, unlike the *AG finetuned* sys-

System	Params	English		Czech	German	Russian
		W&I+L	CoNLL 14	AKCES-GEC	Falko-Merlin	RULEC-GEC
Boyd (2018)	–	–	–	–	45.22	–
Choe et al. (2019)	–	63.05	–	–	–	–
Lichtarge et al. (2019)	–	–	56.8	–	–	
Lichtarge et al. (2020)	–	66.5	62.1	–	–	–
Omelianchuk et al. (2020)	–	72.4	65.3	–	–	–
Rothe et al. (2021) *base*	580M	60.2	54.10	71.88	69.21	26.24
Rothe et al. (2021) *xxl*	13B	69.83	65.65	83.15	75.96	51.62
Rozovskaya and Roth (2019)	–	–	–	–	–	21.00
Xu et al. (2019)	–	63.94	60.90	–	–	–
AG finetuned	210M	69.00	63.40	80.17	73.71	50.20

Table 6: Comparison of selected single-model systems on English (W&I+L, CoNLL-2014), Czech (AKCES-GEC), German (Falko-Merlin GEC) and Russian (RULEC-GEC) datasets. Our reimplementation of the *AG finetuned* model is from Náplava and Straka (2019). Note that models vastly differ in training/fine-tuning data and size (e.g., Rothe et al. (2021) *xxl* is 50 times larger than *AG finetuned*).

System	$M_{0.5}^2$-score					Mean human score				
	NF	NWI	R	SL	Σ	NF	NWI	R	SL	Σ
Original	—	—	—	—	—	8.47	7.99	7.76	7.18	7.61
Korektor	28.99	31.51	46.77	55.93	45.09	8.26	7.60	7.90	7.55	7.63
Synthetic trained	46.83	38.63	46.36	62.20	53.07	8.55	7.99	8.10	7.88	7.98
AG finetuned	65.77	55.20	69.71	71.41	68.08	8.97	8.22	**8.91**	8.35	8.38
GECCC finetuned	**72.50**	**71.09**	**72.23**	**73.21**	**72.96**	**9.19**	**8.72**	**8.91**	**8.67**	**8.74**
Joint GEC+NMT	68.14	66.64	65.21	70.43	67.40	9.06	8.37	8.69	8.19	8.35
Reference	—	—	—	—	—	9.58	9.48	9.60	9.63	9.57

Table 7: Mean score of human judgements and $M_{0.5}^2$ score for each system in domains (NF = *Natives Formal*, NWI = *Natives Web Informal*, R = *Romani*, SL = *Second Learners*, Σ = whole dataset). All results in the whole dataset (the Σ column) are statistically significant with p-value < 0.001, except for the *AG finetuned* and *Joint GEC+NMT* systems, where the p-value is less than 6.2% for $M_{0.5}^2$ score and less than 4.3% for human score, using the Monte Carlo permutation test with 10M samples and probability of error at most 10^{-6} (Fay and Follmann, 2002; Gandy, 2009).

tem. The statistics of the English noise were estimated on *NUCLE* (Dahlmeier et al., 2013), *FCE* (Yannakoudakis et al., 2011) and *W&I+LOCNESS* (Yannakoudakis et al., 2018; Granger, 1998).

4.2 Results and Analysis

Table 7 summarizes the evaluation of the five grammar error correction systems (described in the previous Section 4.1), evaluated with highest-correlating and widely used metric, the M^2 score with $\beta = 0.5$, denoted as $M_{0.5}^2$ (left); and with human judgements (right). For the meta-evaluation of GEC metrics against human judgements, see the following Section 5.

Clearly, learning on GEC annotated data improves performance significantly, as evidenced by

a giant leap between the systems without GEC data (*Korektor*, *Synthetic trained*) and the systems trained on GEC data (*AG finetuned*, *GECCC finetuned* and *Joint GEC+NMT*). Further addition of GEC data volume and domains is statistically significantly better ($p < 0.001$), as the only difference between *AG finetuned* and *GECCC finetuned* systems is that the former uses the *AKCES-GEC* corpus, while the latter is trained on larger and domain-richer *GECCC*. Access to larger data and more domains in the multi-task setting is useful (compare *Joint GEC+NMT* and *AG finetuned* on newly added *Natives Formal* and *Natives Web Informal* domains), although direct training seems superior (*GECCC finetuned* over *Joint GEC+NMT*).

Error Type	#	P	R	$F_{0.5}$
DIACR	3 617	86.84	88.77	87.22
MORPH	610	73.58	55.91	69.20
ORTH:CASING	1 058	81.60	55.15	74.46
ORTH:WSPACE	385	64.44	74.36	66.21
OTHER	3 719	23.59	20.04	22.78
POS	2 735	56.50	22.12	43.10
POS:INFL	1 276	74.47	48.22	67.16
PUNCT	4 709	71.42	61.17	69.10
QUOTATION	223	89.44	61.06	81.83
SPELL	1 816	77.27	75.76	76.96
WO	662	60.00	29.89	49.94

Table 8: Analysis of *GECCC finetuned* model performance on individual error types. For this analysis, all POS-error types were merged into a single error type POS.

We further analyse the best model (*GECCC finetuned*) and inspect its performance with respect to individual error types. For simpler analysis, we grouped all POS-related errors into two error types: *POS* and *POS:INFL* for words which are erroneous only in inflection and share the same lemma with their correction.

As we can see in Table 8, the model is very good at correcting local errors in diacritics (*DIACR*), quotation (*QUOTATION*), spelling (*SPELL*) and casing (*ORTH:CASING*). Unsurprisingly, small changes are easier than longer edits: similarly, the system is better in inflection corrections (*POS:INFL*, words with the same lemma) than on *POS* (correction involves finding a word with a different lemma).

Should the word be split or joined with an adjacent word, the model does so with a relatively high success rate (*ORTH:WSPACE*). The model is also able to correctly reorder words (*WO*), but here its recall is rather low. The model performs worst on errors categorized as *OTHER*, which includes edits that often require rewriting larger pieces of text. Generally, the model has higher precision than recall, which suits the needs of standard GEC, where proposing a bad correction for a good text is worse than being inert to an existing error.

5   Meta-evaluation of Metrics

There are several automatic metrics used for evaluating system performance on GEC dataset, although it is not clear which of them is preferable in terms of high correlation with human judgements on our dataset.

The most popular GEC metrics are the Max-Match (M^2) scorer (Dahlmeier and Ng, 2012) and the ERRANT scorer (Bryant et al., 2017).

The MaxMatch (M^2) scorer reports the F-score over the optimal phrasal alignment between a source sentence and a system hypothesis reaching the highest overlap with the gold standard annotation. It was used as the official metric for the CoNLL 2013 and 2014 Shared Tasks (Ng et al., 2013, 2014) and is also used on various other datasets such as the German *Falko-MERLIN GEC* (Boyd, 2018) or Russian *RULEC-GEC* (Rozovskaya and Roth, 2019).

The ERRANT scorer was used as the official metric of the recent Building Educational Application 2019 Shared Task on GEC (Bryant et al., 2019). The ERRANT scorer also contains a set of rules operating over a set of linguistic annotations to construct the alignment and extract individual edits.

Other popular automatic metrics are the General Language Evaluation Understanding (GLEU) metric (Napoles et al., 2015), that additionally measures text fluency, and I-Measure (Felice and Briscoe, 2015), that calculates weighted accuracy of both error detection and correction.

5.1   Human Judgements Annotation

In order to evaluate the correlation of several GEC metrics with human judgements, we collected annotations of the original erroneous sentences, the manually corrected gold references and automatic corrections made by five GEC systems described in Section 4. We used the hybrid *partial ranking with scalars* (Sakaguchi and Van Durme, 2018), in which the annotators judged the sentences on a scale 0–10 (from ungrammatical to correct).[12] The sentences were evaluated with respect to the context of the document. In total, three annotators judged 1 100 documents, sampled from the test set comprising about 4 300 original sentences and about 15 500 unique corrected variants and gold references of the sentences. The annotators annotated 127 documents jointly and the rest was annotated by a single annotator. This annotation process took about 170 hours. Together with the

---

[12]Recent works (Sakaguchi and Van Durme, 2018; Novikova et al., 2018) both found partial ranking with scalars to be more reliable than direct assessment framework used by WMT (Bojar et al., 2016) and earlier GEC evaluation approaches (Grundkiewicz et al., 2015; Napoles et al., 2015).

Domain	Sentence level		System level	
	r	ρ	r	ρ
Natives Formal	87.13	88.76	92.01	92.52
Natives Web Inf.	80.23	81.47	95.33	91.80
Romani	86.57	86.57	88.73	85.90
Second Learners	78.50	79.97	96.50	97.23
Whole Dataset	79.07	80.40	96.11	95.54

Table 9: Human judgements agreement: Pearson (r) and Spearman (ρ) mean correlation between 3 human judgements of 5 sentence versions at sentence- and system-level.

model training, data preparation and management of the annotation process, our rough estimation is about 300+ man-hours for the correlation analysis per corpus (language).

5.2 Agreement in Human Judgements

For the agreement in human judgements, we report the *Pearson correlation* and *Spearman's rank correlation coefficient* between 3 human judgements of 5 automatic sentence corrections at the system- and sentence-level. At the *sentence level*, the correlation of the judgements about the 5 sentence corrections is calculated for each sentence and each pair of the three annotators. The final sentence-level annotator agreement is the mean of these values over all sentences.

At the *system level*, the annotators' judgements for each system are averaged over the sentences, and the correlation of these averaged judgements is computed for each pair of the three annotators. In order to obtain smoother estimates (especially for Spearman's ρ), we utilize bootstrap resampling with 100 samples of a test set.

The human judgements agreement across domains is shown in Table 9. On the sentence level, the human judgements correlation is high on the least erroneous domain *Natives Formal*, implying that it is easier to judge the corrections in a low error density setting, and it is more difficult in high error density domains, such as *Romani* and *Second Learners* (compare error rates in Table 5).

5.3 Metrics Correlations with Judgements

Following Napoles et al. (2019), we provide a meta-evaluation of the following common GEC metrics robustness on our corpus:

- MaxMatch (M^2) (Dahlmeier and Ng, 2012)
- ERRANT (Bryant et al., 2017)

- GLEU (Napoles et al., 2015)
- I-measure (Felice and Briscoe, 2015)

Moreover, we vary the proportion of recall and precision, ranging from 0 to 2.0 for M^2-scorer and ERRANT, as Grundkiewicz et al. (2015) report that the standard choice of considering precision two times as important as recall may be suboptimal.

While we considered both sentence-level and system-level evaluation in Section 5.2, the automatic metrics should by design be used on a whole corpus, leaving us with only system-level evaluation. Given that the GEC systems perform differently on the individual domains (as indicated by Table 7), we perform the correlation computation on each domain separately and report the average.

For a given domain and metric, we compute the correlation between the automatic metric evaluations of the five systems on one side and the (average of) human judgements on the other side. In order to obtain a smoother estimate of Spearman's ρ and also to estimate standard deviations, we employ bootstrap resampling again, with 100 samples.

The results are presented in Table 10. While Spearman's ρ has more straightforward interpretation, it also has a much higher variance, because it harshly penalizes the differences in the ranking of systems with similar performance (namely *AG finetuned* and *Joint GEC+NMT* in our case). This fact has previously been observed by Macháček and Bojar (2013).

Therefore, we choose the most suitable GEC metric for our *GECCC* dataset according to Pearson r, which implies that $M^2_{0.5}$ and $ERRANT_{0.5}$ are the metrics most correlating with human judgements. Of those two, we prefer the $M^2_{0.5}$ score, not due to its marginal superiority in correlation (Table 10), but rather because it is much more language-agnostic compared to ERRANT, which requires a POS tagger, lemmatizer, morphological dictionary and language-specific rules.

Our results confirm that both M^2-scorer and ERRANT with $\beta = 0.5$ (chosen only by intuition for the CoNLL 2014 Shared task; Ng et al., 2014) correlate much better with human judgements, compared to $\beta = 0.2$ and $\beta = 1$. The detailed plots of correlations of M^2_β score and $ERRANT_\beta$ score with human judgements for β ranging between 0 and 2, presented in Figure 2,
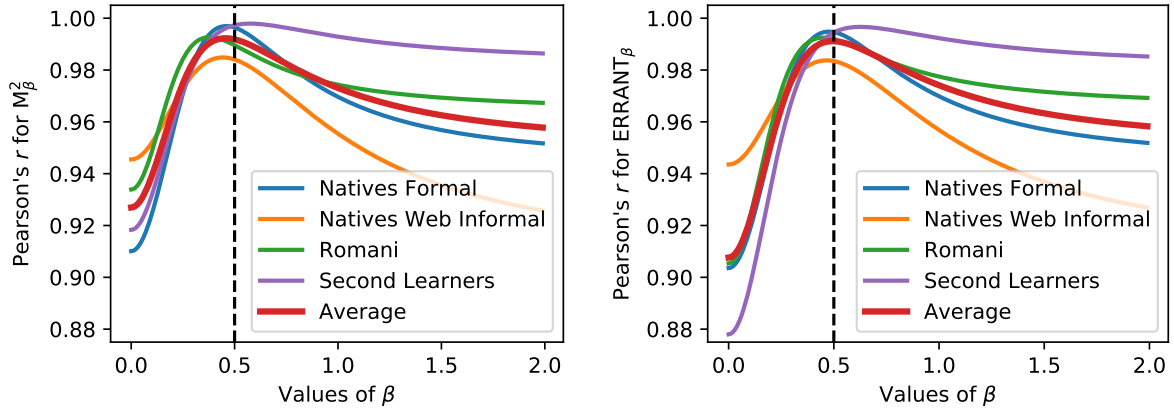
Figure 2: **Left:** System-level Pearson correlation coefficient r between human annotation and M^2_β-scorer for various values of β. **Right:** The same correlation for ERRANT$_\beta$.

Metric	System level	
	r	ρ
GLEU	97.37 ± 1.52	92.28 ± 6.19
I-measure	95.37 ± 2.16	98.66 ± 3.21
$M^2_{0.2}$	96.25 ± 1.71	93.27 ± 9.45
$M^2_{0.5}$	98.28 ± 1.03	97.77 ± 4.27
$M^2_{1.0}$	95.62 ± 1.81	93.22 ± 4.30
ERRANT$_{0.2}$	94.66 ± 2.44	91.19 ± 4.76
ERRANT$_{0.5}$	98.28 ± 1.04	98.35 ± 4.81
ERRANT$_{1.0}$	95.70 ± 1.80	93.61 ± 4.47

Table 10: System-level Pearson (r) and Spearman (ρ) correlation between the automatic metric scores and human annotations.

show that optimal β in our case lies between 0.4 and 0.5. However, we opt to employ the widely used $\beta = 0.5$ because of its prevalence and because the difference to the optimal β is marginal.

Our results are distinct from the results of Grundkiewicz et al. (2015), where $\beta = 0.18$ correlates best on the *CoNLL 14 test set*. Nevertheless, Napoles et al. (2019) demonstrate that $\beta = 0.5$ correlates slightly better than $\beta = 0.2$ on the *FCE* dataset, but that $\beta = 0.2$ correlates substantially better than $\beta = 0.5$ on *Wikipedia* and also on *Yahoo* discussions (a dataset containing paragraphs of Yahoo! Answers, which are informal user answers to other users' questions).

In the latter work, Napoles et al. (2019) propose that larger $\beta = 0.5$ correlate better on datasets with higher error rate and vice versa, given that the *FCE* dataset has 20.2% token error rate, compared

to the error rates of 9.8% and 10.5% of *Wikipedia* and *Yahoo*, respectively. The hypothesis seems to extend to our results and the results of Grundkiewicz et al. (2015), considering that the *GECCC* dataset and the *CoNLL 14 test set* have token error rates of 18.2% and 8.2%, respectively.

5.4 GEC Systems Results

Table 7 presents both human scores for the GEC systems described in Section 4 and also results obtained by the chosen $M^2_{0.5}$ metric. The results are presented both on the individual domains and the entire dataset. Measuring over the entire dataset, human judgements and the M^2-scorer rank the systems in accordance.

Judged by the human annotators, all systems are better than the "do nothing" baseline (the *Original*) measured over the entire dataset, although *Korektor* makes harmful changes in two domains: *Natives Formal* and *Natives Web Informal*. These two domains contain frequent named entities, which upon an eager change disturb the meaning of a sentence, leading to severe penalization by human annotators. *Korektor* is also not capable of deleting, inserting, splitting or joining tokens. The fact that *Korektor* sometimes performs detrimental changes cannot be revealed by the M^2-scorer as it assigns zero score to the *Original* baseline and does not allow negative scores.

The human judgements confirm that there is still a large gap between the optimal *Reference* score and the best performing models. Regarding the domains, the neural models in the finetuned mode that had access to data from all domains seemed

to improve the results consistently across each domain. However, given the fact that the source sentences in the *Second Learners* domain received the worst scores by human annotators, this domain seems to hold the greatest potential for future improvements.

6 Conclusions

We release a new Czech GEC corpus, the *Grammar Error Correction Corpus for Czech* (*GECCC*). This large corpus with 83 058 sentences covers four diverse domains, including essays written by native students, informal website texts, essays written by Romani ethnic minority children and teenagers and essays written by non-native speakers. All domains are professionally annotated for GEC errors in a unified manner, and errors were automatically categorized with a Czech-specific version of ERRANT released at `https://github.com/ufal/errant_czech`. We compare several strong Czech GEC systems, and finally, we provide a meta-evaluation of common GEC metrics across domains in our data. We conclude that M^2 and ERRANT scores with $\beta = 0.5$ are the measures most correlating with human judgements on our dataset, and we choose the $M_{0.5}^2$ as the preferred metric for the *GECCC* dataset. The corpus is publicly available under the CC BY-SA 4.0 license at `http://hdl.handle.net/11234/1-4639`.

Acknowledgements

References

Leonard E Baum and Ted Petrie. 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The annals of mathematical statistics*, 37(6):1554–1563.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Máša Bořkovcová. 2007. *Romský etnolekt češtiny*. Signeta, Praha.

Máša Bořkovcová. 2017. Romský etnolekt češtiny. In Petr Karlík, Marek Nekula, and Jana Pleskalová, editors, *Nový encyklopedický slovník češtiny*. Nakladatelství Lidové Noviny.

Adriane Boyd. 2018. Using Wikipedia Edits in Low Resource Grammatical Error Correction. In *Proceedings of the 4th Workshop on Noisy User-generated Text*. Association for Computational Linguistics.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner Language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Christopher Bryant and Hwee Tou Ng. 2015. How Far are We from Fully Automatic High Quality Grammatical Error Correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, Beijing, China. Association for Computational Linguistics.

Yo Joong Choe, Jiyeon Ham, Kyubyong Park, and Yeoil Yoon. 2019. A Neural Grammatical Error Correction System Built On Better Pre-training and Sequential Transfer Learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–227, Florence, Italy. Association for Computational Linguistics.

Shamil Chollampatt, Weiqi Wang, and Hwee Tou Ng. 2019. Cross-Sentence Grammatical Error Correction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 435–445.

Teodor-Mihai Cotet, Stefan Ruseti, and Mihai Dascalu. 2020. Neural Grammatical Error Correction for Romanian. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 625–631.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better Evaluation for Grammatical Error Correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Vidas Daudaravicius, Rafael E. Banchs, Elena Volodina, and Courtney Napoles. 2016. A Report on the Automatic Evaluation of Scientific Writing Shared Task. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 53–62, San Diego, CA. Association for Computational Linguistics.

Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H. Sanchez Gutierrez, and Kenji Sagae. 2020. Developing NLP Tools with a New Corpus of Learner Spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7238–7243, Marseille, France. European Language Resources Association.

Michael P Fay and Dean A Follmann. 2002. Designing Monte Carlo Implementations of Permutation or Bootstrap Hypothesis Tests. *The American Statistician*, 56(1):63–70.

Mariano Felice and Ted Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 578–587, Denver, Colorado. Association for Computational Linguistics.

Simon Flachs, Ophélie Lacroix, Helen Yannakoudakis, Marek Rei, and Anders Søgaard. 2020. Grammatical Error Correction in Low Error Density Domains: A New Benchmark and Analyses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8467–8478, Online. Association for Computational Linguistics.

G David Forney. 1973. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Axel Gandy. 2009. Sequential Implementation of Monte Carlo Tests With Uniformly Bounded Resampling Risk. *Journal of the American Statistical Association*, 104(488):1504–1511.

Sylvianne Granger. 1998. *Learner English on Computer*, chapter The computer learner corpus: A versatile new source of data for SLA research. Addison Wesley Longman, London & New York.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2019. Minimally-Augmented Grammatical Error Correction. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 357–363, Hong Kong, China. Association for Computational Linguistics.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human Evaluation of Grammatical Error Correction Systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.

Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. 2013a. Facebook Data for Sentiment Analysis. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. 2013b. Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–74, Atlanta, Georgia. Association for Computational Linguistics.

Jan Hajič, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, and Barbora Štěpánková. 2020. MorfFlex CZ 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Maarten Janssen. 2016. TEITOK: Text-Faithful Annotated Corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4037–4043, Paris, France. European Language Resources Association (ELRA).

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords. *CoRR*, abs/2007.03006v1.

Jared Lichtarge, Chris Alberti, and Shankar Kumar. 2020. Data Weighted Training Strategies for Grammatical Error Correction. *Transactions of the Association for Computational Linguistics*, 8:634–646.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora Generation for Grammatical Error Correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3291–3301, Minneapolis, Minnesota. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Jakub Náplava, Martin Popel, Milan Straka, and Jana Straková. 2021. Understanding Model Robustness to User-generated Noisy Texts. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 340–350, Online. Association for Computational Linguistics.

Jakub Náplava and Milan Straka. 2019. Grammatical Error Correction in Low-Resource Scenarios. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 346–356, Stroudsburg, PA, USA. Association for Computational Linguistics.

Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. Enabling Robust Grammatical Error Correction in New Domains: Data Sets, Metrics, and Analyses. *Transactions of the Association for Computational Linguistics*, 7:551–566.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground Truth for Grammatical Error Correction Metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and*

the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 588–593, Beijing, China. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

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 the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

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 the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable Human Ratings for Natural Language Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhan-skyi. 2020. GECToR – Grammatical Error Correction: Tag, Not Rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA, Online. Association for Computational Linguistics.

Michal Richter, Pavel Straňák, and Alexandr Rosen. 2012. Korektor – A System for Contextual Spell-Checking and Diacritics Completion. In *Proceedings of COLING 2012: Posters*, pages 1019–1028, Mumbai, India. The COLING 2012 Organizing Committee.

Alexandr Rosen, Jiří Hana, Barbora Hladká, Tomáš Jelínek, Svatava Škodová, and Barbora Štindlová. 2020. *Compiling and annotating a learner corpus for a morphologically rich language – CzeSL, a corpus of non-native Czech*. Karolinum, Charles University Press, Praha.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A Simple Recipe for Multilingual Grammatical Error Correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL Errors: Challenges and Rewards. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36, Los Angeles, California. 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.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient Online Scalar Annotation with

Bounded Support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.

Karel Šebesta. 2010. Korpusy češtiny a osvojování jazyka. *Studie z aplikované lingvistiky/Studies in Applied Linguistics*, 1:11–34.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Oleksiy Syvokon and Olena Nahorna. 2021. UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language. *CoRR*, abs/2103.16997v1.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and Aspect Error Correction for ESL Learners Using Global Context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shuyao Xu, Jiehao Zhang, Jin Chen, and Long Qin. 2019. Erroneous data generation for Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 149–158, Florence, Italy. Association for Computational Linguistics.

Helen Yannakoudakis, Øistein Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Zheng Yuan and Christopher Bryant. 2021. Document-level Grammatical Error Correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 75–84.