MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection

Elena Volodina[1], **Christopher Bryant**[2],
Andrew Caines[2], **Orphée De Clercq**[3],
Jennifer-Carmen Frey[4], **Elizaveta Ershova**[5], **Alexandr Rosen**[6], **Olga Vinogradova**[7]

[1]University of Gothenburg, Sweden, elena.volodina@svenska.gu.se
[2]ALTA Institute, University of Cambridge, UK, {cjb255,apc38}@cam.ac.uk
[3]LT3, Ghent University, Belgium, orphee.declercq@ugent.be
[4]EURAC Research, Italy, JenniferCarmen.Frey@eurac.edu
[5]JetBrains, Cyprus, elizaveta.ershova@jetbrains.com
[6]Charles University, Czech Republic, alexandr.rosen@ff.cuni.cz
[7]Independent researcher, Israel, olgavinogr@gmail.com

Abstract

This paper reports on the NLP4CALL shared task on Multilingual Grammatical Error Detection (MultiGED-2023), which included five languages: Czech, English, German, Italian and Swedish. It is the first shared task organized by the *Computational SLA*[1] working group, whose aim is to promote less represented languages in the fields of Grammatical Error Detection and Correction, and other related fields. The MultiGED datasets have been produced based on second language (L2) learner corpora for each particular language. In this paper we introduce the task as a whole, elaborate on the dataset generation process and the design choices made to obtain MultiGED datasets, provide details of the evaluation metrics and CodaLab setup. We further briefly describe the systems used by participants and report the results.

1 Introduction

Shared tasks are competitions that challenge researchers around the world to solve practical research problems in controlled conditions (e.g., Nissim et al., 2017; Parra Escartín et al., 2017). Within the field of (second) language acquisition and linguistic issues related to language learning, there have now been several shared tasks on various topics, including:

- argumentative essay analysis for feedback generation[2] (e.g., Picou et al., 2021), where the challenge was to classify text sections into argumentative discourse elements, such as claim, rebuttal, evidence, etc.;
- essay grading / proficiency level prediction (e.g., Ballier et al., 2020), where, given an essay, the major task was to assign a corresponding CEFR proficiency level (A1, A2, B1, B2, etc);
- second language acquisition modeling (e.g., Settles et al., 2018), where the challenge was to predict where a learner might make an error given their error history;

Most prominent, though, have been challenges on so-called grammatical error detection (GED) and correction (GEC), where the task has been to either detect tokens in need of correction, or to produce a correction. Note that the attribute *grammatical* is used traditionally rather than descriptively, since other types of errors (e.g. lexical, orthographical, syntactical) are also targeted. GEC and GED have complemented each other over the years, and the historical interest in the two tasks is visualized in Figure 1. In their comprehensive overview of approaches to GEC, Bryant et al.

[1]The acronym SLA stands for Second Language Acquisition. More information on the working group can be found here: https://spraakbanken.gu.se/en/compsla

[2]https://www.kaggle.com/competitions/feedback-prize-2021/

Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen and Olga Vinogradova. MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection . *Proceedings of the 12th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2023)*. Linköping Electronic Conference Proceedings 197: 1–16.

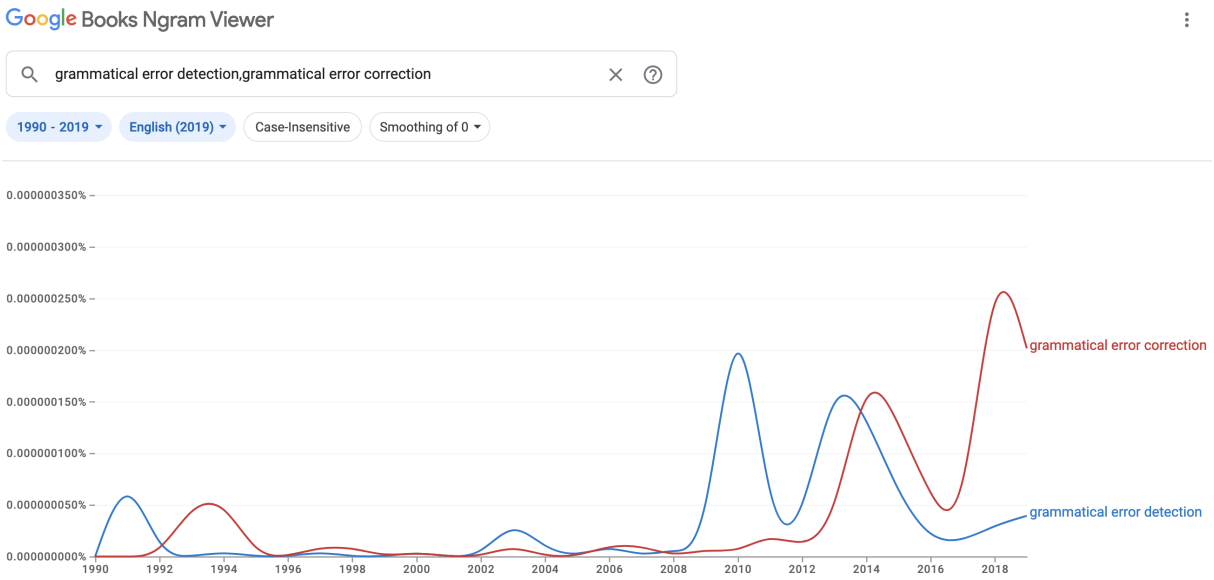

Figure 1: Terms *grammatical error detection* and *grammatical error correction* in Google N-grams (1990–2019)

(2023) observe that most GEC shared tasks have focused only on English, including HOO-2011/12 (Dale and Kilgarriff, 2011; Dale et al., 2012), CoNLL-2013/14 (Ng et al., 2013, 2014), AESW-2016 (Daudaravicius et al., 2016) and BEA-2019 (Bryant et al., 2019), with only a few exploring other languages, such as QALB-2014 and QALB-2015 for Arabic (Mohit et al., 2014; Rozovskaya et al., 2015) and NLPTEA 2014–2020 (Rao et al., 2020) and NLPCC-2018 (Zhao et al., 2018) for Mandarin Chinese.

Though datasets do exist for languages other than English – including for GEC and GED tasks – these rarely feature in shared tasks[3]. Examples of such GEC/GED initiatives are Náplava and Straka (2019) for Czech, Rozovskaya and Roth (2019) for Russian, Davidson et al. (2020) for Spanish, Syvokon and Nahorna (2022) for Ukranian, Cotet et al. (2020) for Romanian, Boyd (2018) for German, Östling and Kurfalı (2022) and Nyberg (2022) for Swedish, to name just a few.

The Matthew effect in GEC and GED? It can be said that the current state of NLP reflects the Matthew effect – i.e., 'the rich get richer, and the poor get poorer' (Perc, 2014; Bol et al., 2018). The Matthew effect has been observed and studied in various disciplines, including economics, sociology, biology, education and even research funding, but is similarly applicable to NLP, as Søgaard (2022) convincingly argued in the article with the provocative title "Should We Ban English NLP for a Year?". The growing bias of NLP research, models and datasets towards English ('the rich') creates inequality by not only making English a 'better equipped language', but also by lowering chances of being cited for researchers working on other languages than English ('the poor'). We witness therefore a tendency in NLP research where reseachers prefer to work on English as it is both the best resourced and best cited language.

To counter-balance the current dynamics in the field towards English dominance, we have taken the initiative to form a *Computational SLA working group* whose main aim is to support and promote work on less represented languages in the area of GED, GEC and other potential tasks in SLA. The MultiGED-2023 shared task is the first one organized by this Computational SLA working group. By bringing non-English datasets, in combination with the English ones, to the attention of the international NLP community, we aim to foster an increasing interest in working on these languages.

2 Task and challenges

The main focus of the first Computational SLA shared task was **error detection**, which we argue should be given more attention as a first step towards pedagogical feedback generation. Through this task, several needs and challenges became clearer which we summarize below.

(i) *Use of authentic L2 data for training al-*

[3]with few exceptions, e.g., UNLP-2023 for Ukranian: https://github.com/asivokon/unlp-2023-shared-task

gorithms. Leacock et al. (2014) convincingly showed that tools for error correction and feedback for foreign language learners benefit from being trained on real L2 students' texts, and that these systems are better suited for use in Intelligent Computer-Assisted Language Learning (ICALL) or Automatic Writing Evaluation (AWE) contexts. Hence the importance of *authentic language learner data*.

(ii) *Focus on less represented languages in GEC/GED*. Both GEC and GED have predominantly been explored in the context of English data. There is a strong incentive to broaden the language spectrum and draw the attention of the international NLP community to other, less represented, languages. We therefore target a few of the less represented languages, namely Czech, German, Italian and Swedish, along with English for comparison with previous work.

(iii) The requirement (i) to use authentic L2 data for the task sets further challenges. First of all, it brings attention to the *scarceness of authentic learner data for a number of languages*. Most languages have modest or tiny collections of L2 data, if any, which contain error annotation and correction. As a consequence, the data is too small to be offered for a shared task by itself. As a way to overcome that problem, we suggest that several languages with smaller datasets coordinate their efforts in a *multilingual low-resource context*, creating possibilities for augmentation of data and/or use of datasets from several languages through domain adaptation, transfer learning, and other modern techniques. The *low-resource context* above refers to a limitation on dataset sizes: there is a maximum of ≈36,000 sentences for each MultiGED language to stimulate creativity in solving problems relating to data scarcity, the smallest datasets comprising ≈8,000 sentences.

(iv) However, (iii) brings further the need to *harmonize datasets* between the languages participating in a multilingual shared task. Harmonization includes both data formatting and data annotation (i.e., converting all language-specific error tags into a set of shared tags). This in itself is a tremendous challenge since languages differ in both linguistic terms and in terms of the annotation approaches and taxonomies adopted by research teams who collated the various corpora. Our initial attempts to convert existing error taxonomies for the five languages to a set of five head categories – punctuation, orthography, lexis, morphology and syntax [POLMS] (Casademont Moner and Volodina, 2022) – proved to be more challenging than expected. As a result, we simplified the task from a multi-class error detection to a binary error detection task, leaving the idea of multi-class detection for future work.

Token	Label	Token	Label
I	c	I	c
saws	i	saws	i
the	c	show	i
show	c	last	c
last	c	nigt	i
nigt	i	.	c
.	c		

Table 1: Data example with two sentences. The sentence on the right demonstrates an error that requires the addition of an extra token, which is indicated by 'i' attached to the next token (see 'i' attached to the token *show* to indicate the missing article *the* before *show*)

MultiGED task in a nutshell The above challenges defined the way the task of *multilingual grammatical error detection in low-resource contexts* was formulated:

> Given an authentic, learner-written sentence, detect tokens within the sentence that contain errors (i.e. perform binary classification on a per-token level) for each provided language separately, or as a multilingual system.

The tokens should be labeled as either correct ('c') or incorrect ('i'), as shown in Table 1.

We encouraged development of multilingual systems that would process all or several languages using a single model, but this was not a mandatory requirement. The submitted systems were evaluated using per-language precision, recall, and $F_{0.5}$ scores. $F_{0.5}$ gives a double weighting to precision over recall, and is conventionally used as the primary metric for GED and GEC on the basis that high precision is more important than high recall for educational applications (Section 4).

The shared task was organized as an open track, in the sense that teams were freely permitted to enhance the provided training and development data for all languages, provided they report the use of additional data, and share them for research

Language	Source corpus	Nr. sentences	Nr. tokens	Nr. errors	Error rate	MultiGED License
Czech	GECCC	35,453	399,742	84,041	0.210	CC BY-SA 4.0
English	FCE	33,243	531,416	50,860	0.096	custom
English	REALEC*	8,136	177,769	16,608	0.093	CC BY-SA 4.0
German	Falko-MERLIN	24,079	381,134	57,897	0.152	CC BY-SA 4.0
Italian	MERLIN	7,949	99,698	14,893	0.149	CC BY-SA 4.0
Swedish	SweLL-gold†	8,553	145,507	27,274	0.187	CC BY-SA 4.0

* We only provide a dev and test set for English-REALEC.

† The original SweLL-gold corpus is released under a CLARIN ID+BY+PRIV+NORED license.

Table 2: MultiGED data statistics.

use and replication studies. This contrasts with a closed track shared task, where teams are prohibited from using additional training and development data beyond that provided by the organizers.

The task aimed to promote research into languages which have received less attention in GED or GEC (Czech, Italian, German, and Swedish alongside English), and for which appropriately annotated datasets are available, even if modest in size (8,000 – 36,000 sentences).

Our **main contributions** are three-fold.

1. We present the first shared task on GED that includes original L2 learner data from Swedish, Italian, German and Czech.

2. We introduce a new dataset of Russian learner English, the REALEC corpus, for the first time.

3. We standardize the formats of several multilingual datasets to faciliate development of multilingual models.

3 Data

We provided training, development and test data for each of the five languages: Czech, English, German, Italian and Swedish.[4] Test sets were released during the test phase through CodaLab and are available there for future work and system comparisons.[5] It is important to note that most corpora are made available on a CC BY-SA 4.0 data license, however the English-FCE uses a custom license, and the original SweLL-gold corpus uses a CLARIN PRIV+ID+BY+NORED license.

3.1 Source data

For each language, a MultiGED dataset was generated from a corpus of original error-annotated learner essays. Table 2 provides an overview of the source corpora, and data statistics of the resulting MultiGED datasets expressed in number of sentences, tokens, errors and error rates. Some of the source corpora mentioned in the Table have already been used in Grammatical Error Detection/Correction research, but we also release two new datasets: one based on REALEC (English) and another on SweLL-gold (Swedish). Where possible, we use the same train/dev/test splits as established in previous work (as is the case for GECCC, FCE, Falko-MERLIN), and only create new splits when necessary (REALEC, Italian MERLIN, SweLL). All datasets were derived from error-annotated L2 learner essays. Below, we provide an overview of each of the source corpora used to create these datasets.

Czech The Grammar Error Correction Corpus for Czech – GECCC (Náplava et al., 2022), consisting of 83,000 sentences, is based on native and non-native texts collected in several earlier projects.[6] The native part consists of essays written by children and teenagers attending primary and secondary schools, either (i) native in standard Czech, or (ii) in its Romani ethnolect, and (iii) informal website texts. However, only the non-native part of GECCC is included in the MultiGED datasets: (iv) essays written by learners of Czech as a foreign or second language, collected mostly for the CzeSL project (Rosen et al., 2020) at nearly all levels of proficiency, from beginners to advanced learners[7] (Rosen et al., 2020),

[4] The training and development splits are available for download on the publicly available MultiGED-2023 github repository: https://github.com/spraakbanken/multiged-2023

[5] https://codalab.lisn.upsaclay.fr/competitions/9784

[6] The corpus is publicly available at http://hdl.handle.net/11234/1-4639

[7] The relatively high share of beginners is the reason why the error rate for Czech in MultiGED is higher than for other languages (Table 2).

but also for the Czech section of MERLIN (Boyd et al., 2014). Instead of relying on the manual and automatic error annotations available in CzeSL and MERLIN, errors in spelling and grammar in the entire GECCC were detected and normalized manually, then categorized automatically using the ERRor ANnotation Toolkit – ERRANT (Bryant et al., 2017), which was modified for Czech.[8] The GECCC corpus is available in its raw untokenized form and in M^2 format (Dahlmeier and Ng, 2012). Basic metadata are available about sex, age and L1 family, with links to a richer set.

English-FCE The FCE Corpus (Yannakoudakis et al., 2011) consists of essays written by candidates for the First Certificate in English (FCE) exam (now "B2 First") designed by Cambridge English to certify learners of English at CEFR level B2. It is part of the larger Cambridge Learner Corpus that has been annotated for grammatical errors (Nicholls, 2003). The FCE Corpus has been used in grammatical error detection (and correction) experiments on numerous occasions, including the BEA 2019 Shared Task (Bryant et al., 2019).

English-REALEC REALEC (Russian Error-Annotated Learner English Corpus) is a corpus of essays written by Russian L1 university students in their final English language examinations designed for students at B1–B2 CEFR levels (Vinogradova and Lyashevskaya, 2022). The requirements for the two types of essays in this examination are the same as in IELTS[9] Task 1 and Task 2. The grammar errors in these essays were annotated manually by specially trained students in the Linguistics Bachelor program. The sentences from all essays were shuffled for the MultiGED shared task to avoid any breach of anonymity, and sentences without any errors identified by the annotators were manually double-checked once more. At both stages of annotating errors and processing sentences for the MultiGED shared task, no stylistic improvements were suggested; all sentences remained authentic.

German For German L2 data, we made use of the Falko-MERLIN GEC corpus as introduced in Boyd (2018). Falko-MERLIN involved the amalgamation of the Falko Corpus – specifically the 248 texts from 'FalkoEssayL2' v2.42 and the 196 texts from 'FalkoEssayWhig' v2.02 (Reznicek et al., 2012) – and 1033 texts from the German section of MERLIN v1.1 (Boyd et al., 2014). Both corpora were annotated in a similar fashion, according to guidelines which demanded only minimal corrections for grammaticality. Falko contains essays at a more advanced proficiency level whereas MERLIN covers a broader range of proficiencies.

Italian The Italian data is drawn from the trilingual learner corpus MERLIN, which contains not only Czech and German texts but also 813 Italian written learner productions (letters and emails), collected within the framework of standardised language tests (Boyd et al., 2014). Similar to the German texts, the handwritten originals of the Italian texts in MERLIN were transcribed and normalised manually, with error annotations added on various levels of linguistic accuracy. Like in the German data, for the shared task we also used the provided minimal corrections for grammaticality, which ignore uncommon stylistic choices.

Swedish For Swedish, we used the SweLL-gold corpus (Volodina et al., 2019), that contains 502 essays written by adult learners at different proficiency levels. The essays were manually transcribed, pseudonymized, normalized and correction annotated. Due to the presence of personal information in the texts, the corpus is under GDPR protection[10] and is distributed for individual use on signing an agreement form. For this reason, texts in their entirety cannot be freely distributed, for example, for use in shared tasks. Shuffling of sentences and removal of demographic information was therefore necessary to make SweLL-gold data openly available for the MultiGED shared task.

3.2 Data pre-processing

The starting point for the corpora featuring in MultiGED varied from dataset to dataset. We took steps to reformat and reshape the corpora so that they were in a common format, as described in Section 3.3 and shown in Table 1. This meant that each corpus needed to be transformed into tabular form with one token per row in the first col-

[8] The modified version of ERRANT, potentially useful for related languages, is available at https://github.com/ufal/errant_czech. However, error tags produced by ERRANT are not used in the MultiGED dataset.

[9] https://www.ielts.org/

[10] https://gdpr-info.eu/

umn and labels in the second column, in line with one of the conventional formats for GED and NLP tasks used more widely. Pre-processing steps for each corpus are described below, starting with the three corpora which have been previously used for GED experiments: Czech GECCC, English FCE and German Falko-MERLIN.

3.2.1 Established GED corpora

For **Czech**, we retained only the learner section of the corpus, which involved first obtaining a list of identifiers for the texts written by L2 learners of Czech (recorded in the 'Domain' field of the metadata file). The GECCC text ID file is aligned with the 'input' file of one sentence per line, but not with the error annotations file (in M^2 format: because M^2 format involves multiple lines per sentence). We therefore attempted to align the original input sentences with the tokenized sentences given in the M^2 file, where tokenization meant that exact matches were often unlikely. We used optimal string alignment as implemented in the `stringdist` package for R (van der Loo, 2014), allowing for a distance up to two-thirds the character length of the original sentence, and breaking any ties manually. Text sequences[11] written by L2 learners were then converted from M^2 to CoNLL format. We used the training, development and test splits already defined in the GECCC.

For the **English-FCE** we started with the M^2 format files made available in the BEA-2019 shared task[12]. The train/dev/test splits are long-established for the FCE Corpus: we simply converted the M^2 files to CoNLL-format and left the splits as they are. To produce files for GED – i.e. with binary error labels – we labelled any token bearing a correction (or following a missing word) as 'i' and all other tokens were labelled 'c'.

Boyd (2018) described the **German** Falko-MERLIN corpus and defined the train/dev/test splits that we use. We obtained the dataset as M^2 files from Adriane Boyd's GitHub repository[13]; note that the data link there carries a security warning and so we made the files available in the German directory of the MultiGED GitHub repository. We converted the M^2 files to CoNLL format[14], and again used the error corrections to arrive at our final token labels, binary 'c' (correct) or 'i' (incorrect).

3.2.2 New GED corpora

Next, we turn to the three corpora which have not previously featured in GED experiments to the best of our knowledge: English REALEC, Italian MERLIN and Swedish SweLL.

Using manually annotated parts of **English REALEC** in .brat format from `https://realec.org/index.xhtml#/exam/`, a tabular representation was produced. Given that the manually annotated subsection of REALEC is relatively small, we only released a development set and a test set for this corpus (i.e., not a training set), randomly assigning each sentence to dev or test. The annotation style in REALEC is different from the other corpora in the shared task: errors are annotated over spans at least one token long. As a result, non-errorful tokens may be included in the span; e.g., [*present-day* rythme → the *present-day* rhythm], which means it is less straightforward to precisely map edit labels to tokens. We nevertheless attempted to automatically infer which tokens should be marked as incorrect using heuristics; e.g. by removing unchanged tokens from the peripheries of both sides of the edit span. Because this conversion process became noisier the longer the error span however, we opted not to attempt it for spans longer than eight tokens, meaning that these longer corrections (just 2.9% of the multiword corrections) are left as they are (i.e. all tokens are labelled as incorrect).

For **Italian MERLIN** we started with the Exmaralda[15] files provided with the 2018 release of the MERLIN corpus (v1.1)[16]. The .exb files contain manually corrected tokenisation and annotations on various layers, including span annotations for error annotation and correction, or token level annotation for edit operations, etc. While the corpus contains annotations for both TH1 (i.e. target hypothesis 1, which only contains form-based corrections of linguistic accuracy) and TH2 (i.e. target hypothesis 2, which also contains meaning-based corrections considering semantics) as de-

[11] Note that not all sequences in the corpora are necessarily *sentences* in a grammatical sense (well-punctuated and containing a finite verb at least), which is why we prefer to refer to them as 'sequences'.

[12] `https://www.cl.cam.ac.uk/research/nl/bea2019st/`

[13] `https://github.com/adrianeboyd/boyd-wnut2018/`

[14] The Python script for this conversion process, `m2_to_conll_conversion-script.py`, is available in the MultiGED repository: `https://github.com/spraakbanken/multiged-2023/`

[15] `https://exmaralda.org/en/`

[16] `http://hdl.handle.net/20.500.12124/6`

fined in Reznicek et al. (2013), we only used the aligned original and TH1 layers of the multilayer annotation.

We transferred the aligned layers into a vertical tab-separated table format, marking any corrections in the normal way as 'i' and uncorrected tokens as 'c'. We omitted lines with unreadable tokens in the original (marked with '-unreadable-' in the token layer), segmented the text where we found sentence-final punctuation in order to insert empty lines between sequences, and applied corrections involving token insertion to the following token in the sequence (in the multilayer annotation of Exmaralda these are indicated against empty tokens). We randomly assigned each sequence to train/dev/test with a probability of .8, .1, .1 respectively.

Finally, for **Swedish** we started with the tabular representation of the data first produced by Casademont Moner and Volodina (2022), which was derived from SweLL-gold in JSON format. As part of processing the corpus, we removed $ symbols (indicating illegible characters), replaced the "-gen" marker with a possessive 's' suffix, and randomly selected one of four options wherever we encountered an anonymisation placeholder. For instance, for any occurrence of the "*-hemland" ('homeland') placeholder, we sampled one of {'Brasil', 'Spanien', 'Irak', 'Kina'} (Brazil, Spain, Iraq, China); and for any occurrence of the "*-svensk-stad" ('Swedish town') placeholder, we sampled a made-up placename from {'Sydden', 'Norrebock', 'Rosaborg', 'Ögglestad'}. Similar fake replacements were made for '*-geoplats' ('geolocation'), '*-plats' ('place'), '*-institution', '*-skola' ('school'), '*-land' ('country'), '*-region', '*-stad' ('town'), '*-linjen' ('transport line').

As a GDPR-related requirement of using SweLL, we randomly shuffled the order of sentences in order to protect individual privacy. We then assigned the sentences to train/dev/test splits with a probability of .8, .1, .1 respectively. As with Italian MERLIN, in SweLL the insertion correction type is marked against an empty token: therefore we carried such annotations forward to the next token, in line with other corpora in MultiGED, and omitted the empty tokens. Subsequently, the usual 'i' and 'c' labels were generated based on the presence of corrections (or not) against each token in the file.

3.3 Data format

MultiGED data is, thus, provided in a tab-separated format consisting of two columns and no headers: the first column contains the token and the second column contains the label (`c` or `i`), as shown in Table 1. Each sequence is separated by an empty line, and double quotes are escaped (`\"`). Error labels (`i`) are attached on the same line where the errors are, with one exception: if an insertion is necessary, the `i` label is attached to the next token; e.g., the right-hand side of Table 1. System outputs should be generated in the same format.

4 Evaluation

System evaluation was carried out in terms of token-based $F_{0.5}$ to be consistent with previous work in error detection (Bell et al., 2019; Kaneko and Komachi, 2019; Yuan et al., 2021). It has been customary to evaluate GED/GEC systems in terms of $F_{0.5}$, which weights precision twice as much as recall, since the CoNLL-2014 shared task, given that it is more important to an end user that a system makes a correct prediction than to necessarily detect all errors (Ng et al., 2014). Precision (P), Recall (R) and F-score (F_β) were hence calculated in the standard way based on the total number of true positives (TP), false positives (FP) and false negatives (FN) (Equation 1–3) with the parameter $\beta = 0.5$.

$$P = \frac{TP}{TP + FP} \quad (1) \qquad R = \frac{TP}{TP + FN} \quad (2)$$

$$F_\beta = (1 + \beta^2) \times \frac{P \times R}{(\beta^2 \times P) + R} \quad (3)$$

One notable limitation of token-based $F_{0.5}$ is that systems will receive multiple rewards for detecting each erroneous token in a multi-word edit, e.g. [*In other hand* → *On the other hand*], when it might otherwise be more realistic to treat such cases as a single error. This approximation is generally acceptible, however, given that multi-token errors are typically much rarer than single token errors, and it may in fact be beneficial to reward systems for the partial detection of multi-token errors. It is nevertheless worth keeping this property of token-based evaluation in mind.

Team	System description
EliCoDe Colla et al. (2023)	XLM-RoBERTa language model pretrained on ≈100 languages with a stacked linear classifier on top, with a dropout layer in-between fine-tuned 5 different models for 5 languages on train (or train+dev) data
DSL-MIM-HUS Ngo et al. (2023)	XLM-RoBERTa language model from the HuggingFace repo pretrained on ≈100 languages, fine-tuned jointly on all MultiGED datasets i.e. there is only one trained model for prediction of all the test datasets
Brainstorm Thinkers	mBERT, for all six datasets
VLP-char (no eng-realec) Ngo et al. (2023)	character-based LSTM model with two recurrent layers, unidirectional supervised approach, separate model for each dataset, REALEC excluded no external datasets
NTNU-TRH Bungum et al. (2023)	multilingual system based on LSTMs, GRUs and standard RNNs with multilingual Flair embeddings for a sequence-to-sequence labeling multitask learning
su-dali (only swe) Kurfalı and Östling (2023)	distantly-supervised transformer-based machine translation (MT) system trained solely on artificial dataset of 200 million sentences, only Swedish no supervision, training or fine-tuning on any labeled data

Table 3: Overview of submitted systems, lister in the order of registration

4.1 CodaLab

Evaluation was formally carried out on the CodaLab competition platform[17], with participants being allowed to anonymously make a maximum of 2 submissions on the test data during the test phase. Each submission was expected to contain output for as many languages as the team wished to participate in, and so participants could effectively make a maximum of 2 submissions for each dataset in the shared task.

It is **extremely important** to note that we treated the best score *from either submission* as the official result for each team. This means that if a team scored 50 in Language A and 60 in Language B from Submission 1, but 45 in Language A and 70 in Language B from Submission 2, the official score for the team is 50 in Language A (Submission 1) and 70 in Language B (Submission 2). In other words, we did not penalise teams for uploading their best system output in different submissions.

5 Teams, Approaches, Results

In total, six teams participated in the task, representing five different countries: China, Italy, Norway, Sweden and Vietnam. Four teams developed systems for all five languages (and six datasets): EliCoDe (Colla et al., 2023), NTNU-TRH (Bungum et al., 2023), DDSL-MIM-HUS (Ngo et al., 2023, System 1) and Brainstorm Thinkers (no submitted system description); one team submitted results for all five languages excluding the English-REALEC dataset: VLP-char (Ngo et al., 2023, System 2); and one team submitted results for Swedish only: su-dali (Kurfalı and Östling, 2023).

The different approaches that each team took are summarized in Table 3. The most successful approaches relied on BERT-like large language models (see Table 4). The team with the best average result across all languages, EliCoDe, fine-tuned a different model for each dataset and showed considerably superior recall capabilities on most datasets (Colla et al., 2023). The second-best average result came from the DSL-MIM-HUS team, who fine-tuned one pre-trained model on all 6 datasets at once (Ngo et al., 2023). The same team also trained a character-based LSTM, VLP-char. The NTNU-TRH team used LSTMs as well, implementing their systems with FlairNLP and comparing monolingual and multilingual scenarios (Bungum et al., 2023). These latter approaches require less data for training but show weaker performance in recall and precision, either tending to detect fewer errors or produce a greater number of false positives. The su-dali team used artificial data mimicking the error distribution from the Swedish source corpus, and achieved very good results on Swedish showing that access to manually annotated training data can be avoided (Kur-

[17] https://codalab.lisn.upsaclay.fr/competitions/9784

a. Results on Czech

Team	P	R	$F_{0.5}\uparrow$
EliCoDe	82.01	51.79	**73.44**
DSL-MIM-HUS	58.31	55.69	57.76
Brainstorm Thinkers	62.35	23.44	46.81
VLP-char	34.93	63.95	38.42
NTNU-TRH	80.65	6.49	24.54
Majority	84.32	43.22	70.85

b. Results on English – FCE

Team	P	R	$F_{0.5}\uparrow$
EliCoDe	73.64	50.34	**67.40**
DSL-MIM-HUS	72.36	37.81	61.18
Brainstorm Thinkers	70.21	37.55	59.81
VLP-char	20.76	29.53	22.07
NTNU-TRH	81.37	1.84	8.45
Majority	85.35	32.48	64.39

c. Results on English – REALEC

Team	P	R	$F_{0.5}\uparrow$
DSL-MIM-HUS	62.81	28.88	**50.86**
EliCoDe	44.32	40.73	43.55
Brainstorm Thinkers	48.19	31.22	43.46
NTNU-TRH	51.34	1.13	5.19
Majority	65.46	27.23	51.11

d. Results on German

Team	P	R	$F_{0.5}\uparrow$
EliCoDe	84.78	73.75	**82.32**
DSL-MIM-HUS	77.80	51.92	70.75
Brainstorm Thinkers	77.94	47.55	69.11
NTNU-TRH	83.56	15.58	44.61
VLP-char	25.18	44.27	27.56
Majority	87.80	49.88	76.21

e. Results on Italian

Team	P	R	$F_{0.5}\uparrow$
EliCoDe	86.67	67.96	**82.15**
DSL-MIM-HUS	75.72	38.67	63.55
Brainstorm Thinkers	70.65	36.46	59.49
NTNU-TRH	93.38	19.84	53.62
VLP-char	25.79	44.24	28.14
Majority	90.25	40.95	72.74

f. Results on Swedish

Team	P	R	$F_{0.5}\uparrow$
EliCoDe	81.80	66.34	**78.16**
DSL-MIM-HUS	74.85	44.92	66.05
Brainstorm Thinkers	73.81	39.94	63.11
su-dali	82.41	27.18	58.60
VLP-char	26.40	55.00	29.46
NTNU-TRH	80.12	5.09	20.31
Majority	89.90	45.37	75.15

Table 4: Results for each language and team in terms of Precision (P), Recall (R) and F-score ($F_{0.5}$). The *Majority* score is based on the majority predicted token-based labels across all systems.

falı and Östling, 2023).

Czech Systems that relied on Transformer-based architectures (the top three in Table 4) achieved the top-3 $F_{0.5}$ scores. Despite that, the best recall comes from the LSTM-based system (VPL-char).

English-FCE The performance of the RoBERTa-based architecture, fine-tuned exclusively on the FCE dataset by EliCoDe team, outperformed other architectures in all evaluation metrics, indicating its superior efficacy for the FCE dataset.

English-REALEC The results obtained from the REALEC dataset were relatively low compared to other datasets, which may be attributed to the different annotation style in REALEC (see Section 3.2), and the fact that REALEC was both released later in the shared task and without a training split.

German The highest scores were obtained by all teams on the German Falko-MERLIN dataset. Remarkably, the teams NTNU-TRH and VLP-char, who did not use external data, exhibited substantially better performance on the German dataset.

Italian The solutions submitted for the German and Italian datasets exhibited the highest performance levels compared to the other datasets. This finding could potentially be attributed to the fact that these datasets were sourced from the MERLIN corpus and possessed a high level of consistency in their annotations.

Swedish The Swedish dataset received the highest participation rate among all the datasets. The best performance was achieved by Transformer-based architectures, which is consistent with the performance on other datasets. Nevertheless, satisfactory results were also achieved by solutions using LSTMs without pre-training or additional data.

Altogether, shared task participants submitted different systems representing a variety of approaches, including machine translation, LSTMs, mBERT and XLM-RoBERTa (Table 3). The best results were achieved by teams employing the multilingual XLM-RoBERTa (large) language model pre-trained on ≈100 languages (Conneau et al., 2020). The systems trained and fine-tuned

Language	Team	Best $F_{0.5}$ ↓
German	EliCoDe	82.32
Italian	EliCoDe	82.15
Swedish	EliCoDe	78.16
Czech	EliCoDe	73.44
Eng-FCE	EliCoDe	67.40
Eng-REALEC	DSL-MIM-HUS	50.87

Table 5: Best results for each language dataset.

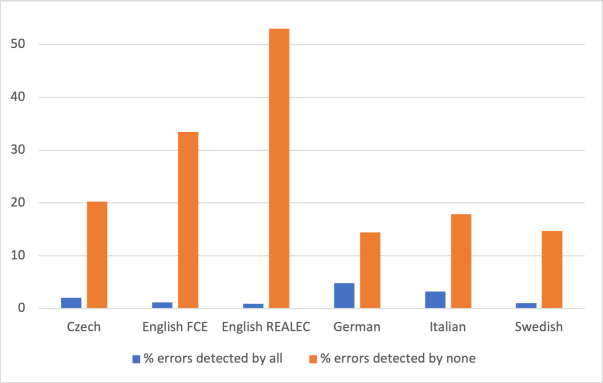

Figure 2: Percentage of errors in the test set which were either detected by all (blue bars, on the left) or none (orange bars, on the right) of the participating teams.

separately for each language dataset by the EliCoDe team performed substantially better than the ones that used one multilingual model for all languages (team DSL-MIM-HUS), with the exception of the English-REALEC dataset, where the results were reversed (see the results for the top-performing systems in Table 5). This is an important insight, because the EliCoDe team also showed that for some language datasets multilingual models, fine-tuned on all datasets, performed better than monolingually fine-tuned ones (Colla et al., 2023). On the one hand, it is intuitive that monolingual models might perform better than multilingual models because they are more specially trained for a particular target language, but on the other hand, multilingual models might be expected to perform better because they have access to richer multilingual representations from linguistically-related languages. In either case, both approaches have different advantages which are worth exploring further.

Table 4 also lists the scores from a token-based majority vote for each language in gray. This is based on the performance of a system relying on a majority vote among all system outputs. For the two languages with an even number of system outputs – English-REALEC and Swedish – a fallback was implemented in case of a tie, namely to choose the output of the best system (EliCoDe in both languages). As can be observed, this majority system led to better precision in all languages and lower recall. If this score were to be included in the ranking, it would end up on place two for all languages, except for English-REALEC where, with an $F_{0.5}$ of 51.11 it would obtain first place.

In Figure 2 we combine all system output to get more insights in the error detection (the `i` labels). The blue bars (on the left) represent the percentage of errors that were detected by all participatings systems in each language, whereas the orange bars (on the right) illustrate the percentage of errors none of which the systems were able to detect. What draws the attention are the high percentages of errors none of the approaches were able to detect for English (33% for English FCE and 53% for English REALEC, respectively). Also, when ranked by best results for all languages (Table 5) it is counter-intuitive to see that English comes at the bottom, as English has typically received the most attention in GED. REALEC is a special case – we did not provide training data for it, and obviously models trained on other languages or other datasets for the same language did not generalize well to REALEC – hypothetically because REALEC had a different type of annotation approach. However, an interesting question is why performance on the English-FCE dataset was lower than on all other languages? In this respect, the EliCoDe team (Colla et al., 2023) carried out an analysis of training/development splits versus the test split per language for linguistic similarity and identified bigger differences between English splits than any other MultiGED languages; they conclude this may be the reason why scores were lower on English.

A short look at the six system output files for Swedish shows that most of the errors that all systems missed (i.e. labeled them as `c` instead of `i`) are those that cover:

- lexical choices, for example non-idiomatic use of vocabulary, e.g. `Jag tror att religion *` **`har`** `ingen roll...`[18] ('I think that religion ***has** no role...')
- verb tense harmonization with other verb

[18] The missed token shown in bold.

tenses used in the sentence, e.g. `Hon tycker att Hans är hennes äkta kärlek men så *var det inte` (‘She thinks that Hans is her real love, but it ***was** not the case’)

- a few preposition and syntactic construction choices, e.g. `Hur går det *med dig?` (‘How is it going ***with** you?’)
- few of the errors missed by all systems would in fact require longer context than one sentence for determiniting the need of a correction

Note that these are only indicative insights and a more thorough analysis would be necessary to draw any proper conclusions.

Rather obviously, spelling errors resulting in ‘non-words’ (OOVs – out-of-vocabulary strings) were easier to detect than errors resulting in some existing word forms (‘real-word errors’). Whereas the entire Czech test data included 6.937% of non-words, there were much fewer non-words among the 1716 incorrect word forms that all the systems failed to detect: 0.047%. The almost 15:1 ratio was lower for the English data (about 7:1 for FCE: 1.440% vs. 0.199%; 4:1 for REALEC: 1.135% vs. 0.310%), but it is still clear that real-word errors were harder to detect.

In future, it would be useful to see error distributions made by systems by types of (gold) error labels [e.g. POLMS[19]] and account for their effect on different language systems performance. Another possible interesting analysis could be to correlate system performance with learners’ language proficiency, their first languages, as well as with the effect of essay tasks on system performance.

6 Comparison with previous work

To provide some context for the MultiGED results on the English FCE benchmark, we present Table 6, which summarisee results on English GED in the past five years. The state-of-the-art has been gradually pushed: Bell et al. (2019) explored the effect of using different contexual embeddings and their generalizability to different datasets, showing the potential of “leveraging information learned in an unsupervised manner from high volumes of unlabeled data” and their sensitivity to error types, with BERT embeddings (Peters et al., 2017) being especially promising ($F_{0.5}$ 57.28). Kaneko and Komachi (2019) complemented $BERT_{BASE}$ with a Multi-Head Multi-Layer Attention (MHMLA) function to achieve a new state of the art for GED, reaching $F_{0.5}$ 61.65 on FCE. Yuan et al. (2021) meanwhile showed that ELECTRA (Clark et al., 2020) has a “discriminative pre-training objective that is conceptually similar to GED”, which improved GED results by a large margin on several public English datasets, reaching $F_{0.5}$ 72.93 on the FCE benchmark. Two years later, the results by Yuan et al. (2021) are still state-of-the-art. The bulk of work on English provides potential ways for improvement on other MultiGED languages – if nothing else, to see whether the same trends hold cross-linguistically.

System / English FCE	P	R	$F_{0.5}$
`MultiGED-23`			
EliCoDe	73.64	50.34	**67.40**
DSL-MIM-HUS	72.36	37.81	61.18
`State-of-the-art`			
Yuan-2021, BERT	75.73	47.98	67.88
Yuan-2021, XLNet	77.50	49.81	69.75
Yuan-2021, ELECTRA	82.05	50.49	**72.93**
`Previous results`			
Kaneko-Komachi-2019	68.87	43.45	61.65
Bell-2019, $BERT_{BASE}$	64.96	38.89	57.28

Table 6: Comparison to previous GED results on English FCE dataset (Yuan et al., 2021; Kaneko and Komachi, 2019; Bell et al., 2019).

We are unable to make similar comparisons for the other languages in MultiGED because this is the first time these languages have been evaluated in the context of GED. More specifically:

- For Czech, previous research explores grammatical error correction (GEC) rather than detection (e.g. Náplava and Straka, 2019; Náplava et al., 2022). There has been some previous work on the evaluation of Czech error detection in the context of a spellchecking tool, Korektor (Ramasamy et al., 2015), however, this is not fully compatible with the scope of errors in MultiGED.
- For German, although there is some work on sentence-level error detection (e.g. Boyd, 2012) and error correction (e.g. Boyd, 2018; Sun et al., 2022; Pająk and Pająk, 2022), there is no previous work on token-level GED.

[19]POLMS = P-unctuation, O-rthography, L-exical, M-orphology, S-yntax

Feedback type	Example	NLP task
1. correct/incorrect	*incorrect*	sentence-level acceptability judgment
2. highlighting	I saw show last night .	GED – grammatical error detection (per token)
3. metalinguistic	*note definiteness / morphology*	multi-class GED
4. error explanation	*note rules for noun definiteness*	instructive feedback generation
5. correct answer	I saw **the** show last night .	GEC – grammatical error correction
6. level/grade	CEFR level A2	AEG – automatic essay grading

Table 7: NLP tasks for different feedback types

- For Italian, we are unaware of any work on GED or GEC at all.
- For Swedish, rule-based error detection was developed within the Granska project, (e.g. Birn, 2000; Arppe, 2000), however, it is difficult to use these results for comparison since the evaluation metrics and test sets are different, as is the scope of errors.

We can therefore conclude that the MultiGED-2023 shared task has established a new set of benchmark datasets and state-of-the-art GED baselines for four new languages in this domain: Czech, German, Italian and Swedish.

7 Concluding remarks

We have presented datasets and results for the task of multilingual grammatical error detection for five languages and six corpora, three of which have not previously featured in the domain of GED.

We view this contribution *primarily* as a step towards empowering "smaller" languages and decreasing the Matthew effect in this field (Søgaard, 2022; Perc, 2014; Bol et al., 2018). It is our hope that the availability of these datasets and baselines will spark further GED research for these languages. *Secondly*, we view this shared task as a step towards instructional feedback generation in ICALL tutoring systems – corrections, error classification and grammar explanations being reserved as potential future shared tasks, see Table 7 for some ideas.

Besides this, we summarise a few of our insights that might be useful to keep in mind for further GED experiments:

1. Pre-trained large language models have no doubt pushed the field far forward (cf. Yuan et al., 2021; Colla et al., 2023; Ngo et al., 2023). It is left to see in the future how GPT[20] models can influence the field (e.g. Radford et al., 2018; Wu et al., 2023; Lund and Wang, 2023).
2. Monolingual fine-tuning tends to outperform multilingual approaches, however, there are some exceptions (Colla et al., 2023; Ngo et al., 2023; Bungum et al., 2023), and more attention should be given to multilingual approaches.
3. Embeddings of various types can have a significant impact on system performance (Bungum et al., 2023).
4. Artificial data containing error distributions similar to the test data facilitates reaching competitive performance with relatively low costs (Kurfalı and Östling, 2023), and is a promising way to go.
5. The quality of data annotation is critical for high performance, as has been indicated by the results on different MultiGED languages, the ones coming from MERLIN (German and Italian) showing better results compared to other annotation paradigms (see Section 5 for descriptions of Italian).

Finally, we would like to encourage those who have L2 data and are willing to use it for a shared task on L2 language *in combination with other languages*, to make contact with the *Computational SLA working group*.[21] It would be especially welcome if languages from beyond the Indo-European group could feature in future shared tasks.

Acknowledgements

The first author has been supported by the Swedish *Språkbanken Text* and by *HumInfra* through funding from the Swedish Research Council (contracts

[20] GPT stands for Generative Pretrained Transformers

[21] https://spraakbanken.gu.se/en/compsla

2017-00626 and 2021-00176). The second and third authors are supported by Cambridge University Press & Assessment.

References

Antti Arppe. 2000. Developing a grammar checker for Swedish. In *Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999)*, pages 13–27, Trondheim, Norway. Department of Linguistics, Norwegian University of Science and Technology, Norway.

Nicolas Ballier, Stéphane Canu, Caroline Petitjean, Gilles Gasso, Carlos Balhana, Theodora Alexopoulou, and Thomas Gaillat. 2020. Machine learning for learner English: A plea for creating learner data challenges. *International Journal of Learner Corpus Research*, 6(1):72–103.

Samuel Bell, Helen Yannakoudakis, and Marek Rei. 2019. Context is key: Grammatical error detection with contextual word representations. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 103–115, Florence, Italy. Association for Computational Linguistics.

Juhani Birn. 2000. Detecting grammar errors with lingsoft's Swedish grammar checker. In *Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999)*, pages 28–40, Trondheim, Norway. Department of Linguistics, Norwegian University of Science and Technology, Norway.

Thijs Bol, Mathijs de Vaan, and Arnout van de Rijt. 2018. The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, 115(19):4887–4890.

Adriane Boyd. 2012. *Detecting and diagnosing grammatical errors for beginning learners of german: From learner corpus annotation to constraint satisfaction problems*. Ph.D. thesis, The Ohio State University.

Adriane Boyd. 2018. Using Wikipedia Edits in Low Resource Grammatical Error Correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium. 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, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *arXiv preprint arXiv:2211.05166*.

Lars Bungum, Björn Gambäck, and Arild Brandrud Næss. 2023. NTNU-TRH System at the MultiGED-2023 Shared Task on Multilingual Grammatical Error Detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL)*.

Judit Casademont Moner and Elena Volodina. 2022. Swedish MuClaGED: A new dataset for Grammatical Error Detection in Swedish. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 36–45.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR*. OpenReview.net.

Davide Colla, Matteo Delsanto, and Elisa Di Nuovo. 2023. ELICODE at MultiGED2023: fine-tuning XLM-RoBERTa for multilingual grammatical error detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada. Association for Computational Linguistics.

Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 Pilot Shared Task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France. 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 Twelfth Language Resources and Evaluation Conference*, pages 7238–7243, Marseille, France. European Language Resources Association.

Masahiro Kaneko and Mamoru Komachi. 2019. Multi-Head Multi-Layer Attention to Deep Language Representations for Grammatical Error Detection. *Computación y Sistemas*, 23(3).

Murathan Kurfalı and Robert Östling. 2023. A distantly supervised Grammatical Error Detection/Correction system for Swedish. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL)*.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool.

Mark P.J. van der Loo. 2014. The stringdist Package for Approximate String Matching. *The R Journal*, 6(1):111–122.

Brady D Lund and Ting Wang. 2023. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News*.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The First QALB Shared Task on Automatic Text Correction for Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47, Doha, Qatar. 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, Hong Kong, China. Association for Computational Linguistics.

Jakub Náplava, Milan Straka, Jana Straková, and Alexandr Rosen. 2022. Czech grammar error correction with a large and diverse corpus. *Transactions of the Association for Computational Linguistics*, 10:452–467.

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.

The Quyen Ngo, Thi Minh Huyen Nguyen, and Phuong Le-Hong. 2023. Two Neural Models for Multilingual Grammatical Error Detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL)*.

Diane Nicholls. 2003. The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference; UCREL technical paper number 16*. Lancaster University.

Malvina Nissim, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling. 2017. Last words: Sharing is caring: The future of shared tasks. *Computational Linguistics*, 43(4):897–904.

Martina Nyberg. 2022. *Grammatical Error Correction for Learners of Swedish as a Second Language*. Master's thesis, Uppsala university.

Krzysztof Pająk and Dominik Pająk. 2022. Multilingual fine-tuning for grammatical error correction. *Expert Systems with Applications*, 200:116948.

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical Considerations in NLP Shared Tasks. In *Proceedings of the First ACL Workshop on Ethics in*

Natural Language Processing, pages 66–73, Valencia, Spain. Association for Computational Linguistics.

Matjaž Perc. 2014. The Matthew effect in empirical data. *Journal of The Royal Society Interface*, 11(98):20140378.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

Aigner Picou, Alex Franklin, Maggie Meg Benner, Perpetual Baffour, Phil Culliton, Ryan Holbrook, Scott Crossley, and Terry_yutian Ulrichboser. 2021. Feedback Prize - Evaluating Student Writing.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Loganathan Ramasamy, Alexandr Rosen, and Pavel Stranák. 2015. Improvements to Korektor: A Case Study with Native and Non-Native Czech. In *Proceedings ITAT 2015: Information Technologies - Applications and Theory*, volume 1422 of *CEUR Workshop Proceedings*, pages 73–80. CEUR-WS.org.

Gaoqi Rao, Erhong Yang, and Baolin Zhang. 2020. Overview of NLPTEA-2020 shared task for Chinese grammatical error diagnosis. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 25–35, Suzhou, China. Association for Computational Linguistics.

Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the falko corpus. *Automatic treatment and analysis of learner corpus data*, 59:101–123.

Marc Reznicek, Anke Lüdeling, Cedric Krummes, and Franziska Schwantuschke. 2012. *Das FalkoHandbuch. Korpusaufbau und Annotationen Version 2.0.*

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.

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

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xin Sun, Tao Ge, Shuming Ma, Jingjing Li, Furu Wei, and Houfeng Wang. 2022. A Unified Strategy for Multilingual Grammatical Error Correction with Pre-trained Cross-Lingual Language Model. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 4367–4374, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization.

Oleksiy Syvokon and Olena Nahorna. 2022. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. *arXiv preprint arXiv:2103.16997*.

Olga Vinogradova and Olga Lyashevskaya. 2022. Review Of Practices Of Collecting And Annotating Texts In The Learner Corpus REALEC. In *Text, Speech, and Dialogue: 25th International Conference, TSD 2022*, page 77–88, Berlin, Heidelberg. Springer-Verlag.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, et al. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology (NEJLT)*, 6:67–104.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark. *arXiv preprint arXiv:2303.13648*.

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, Shiva Taslimipoor, Christopher Davis, and Christopher Bryant. 2021. Multi-Class Grammatical Error Detection for Correction: A Tale of

Two Systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8722–8736, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the NLPCC 2018 Shared Task: Grammatical Error Correction. In *Natural Language Processing and Chinese Computing*, pages 439–445. Springer International Publishing.

Robert Östling and Murathan Kurfalı. 2022. Really good grammatical error correction, and how to evaluate it. *Proceedings of Swedish Language Technology Conference*.