

Towards better language representation in Natural Language Processing

A multilingual dataset for text-level Grammatical Error Correction

Arianna Masciolini,^{1,2} Andrew Caines,³ Orphée De Clercq,⁴
Joni Kruijsbergen,⁴ Murathan Kurfali,⁵
Ricardo Muñoz Sánchez,^{1,2} Elena Volodina,^{1,2}
Robert Östling,⁶ Kais Allkivi,⁷ Špela Arhar Holdt,⁸
Ilze Auzina,⁹ Roberts Dargis,⁹ Elena Drakonaki,¹⁰
Jennifer-Carmen Frey,¹¹ Isidora Glišić,¹² Pinelopi Kikilintza,¹⁰
Lionel Nicolas,¹¹ Mariana Romanyshyn,¹³ Alexandr Rosen,¹⁴
Alla Rozovskaya,¹⁵ Kristjan Suluste,¹⁶ Oleksiy Syvokon,¹⁷
Alexandros Tantos,¹⁰ Despoina-Ourania Touriki,¹⁰
Konstantinos Tsiotskas,¹⁰ Eleni Tsourilla,¹⁰
Vassilis Varsamopoulos,¹⁰ Katrin Wisniewski,¹⁸ Aleš Žagar⁸
and Torsten Zesch¹⁹

¹ Språkbanken Text, SFS, University of Gothenburg | ² University of
Cambridge | ³ Ghent University | ⁴ RISE Research Institutes of Sweden |
⁵ Stockholm University | ⁶ Tallinn University | ⁷ University of Ljubljana |
⁸ University of Latvia | ⁹ Aristotle University of Thessaloniki | ¹⁰ Eurac
Research Bolzano | ¹¹ University of Iceland | ¹² Grammarly | ¹³ Charles
University | ¹⁴ City University of New York | ¹⁵ Institute of the Estonian
Language | ¹⁶ Microsoft | ¹⁷ Leipzig University | ¹⁸ FernUniversität in
Hagen

This paper introduces MultiGEC, a dataset for multilingual Grammatical Error Correction (GEC) in twelve European languages: Czech, English, Estonian, German, Greek, Icelandic, Italian, Latvian, Russian, Slovene, Swedish and Ukrainian. MultiGEC distinguishes itself from previous GEC datasets in that it covers several underrepresented languages, which we argue should be included in resources used to train models for Natural Language Processing tasks which, as GEC itself, have implications for Learner Corpus Research and Second Language Acquisition. Aside from multilingualism, the novelty of the MultiGEC dataset is that it consists of full texts – typically learner essays – rather than individual sentences, making it possible to train systems that take a broader context into account. The dataset was built for MultiGEC-2025, the first shared task in multilingual text-level GEC, but it remains accessible after its competitive phase, serving as a resource to train new error correction systems and perform cross-lingual GEC studies.

Keywords: learner corpora, grammatical error correction, multilingual corpora, Matthew effect, MultiGEC shared task

1. Introduction

There is a tendency in Natural Language Processing (NLP) research to work on English as it is both the better resourced and better cited language (e.g. Søgaard, 2022). This dynamic is consistent with the *Matthew effect* (Perc, 2014), which derives its name from the Gospel of Matthew and denotes the principle of “the rich getting richer and the poor getting poorer”¹. The term *Matthew effect* was coined by Merton (1968) to describe an “accumulative advantage” in scientific recognition and has since been applied to describe the effect where some selected people/areas (“the rich”) inadvertently dominate over others (“the poor”).

While Merton (1968) studied this effect in recognition of achievements in the field of science, it is equally applicable to other areas. There is a trend in terms of distribution of research funding where those who already have received funding tend to get more funding in the next rounds (e.g. Bol et al., 2018). We can also see this effect through the availability of digital language data and the research carried out for GEC, as illustrated in Figure 1, which shows the number of GEC-related

1. Named as such due to the Gospel of Matthew 25:29 – ‘For to every one who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away.’

publications found for each of the twelve languages featured in our multilingual dataset, and discussed in further depth by Masciolini et al. (2025b).

At least within NLP research, English has become an unintentional beneficiary of the *Matthew effect* (getting the status of “the rich”). Clearly, several factors have contributed to this state, among others: 1. the availability of English data that can be used for training algorithms; 2. the status of English as the current scientific *lingua franca* – the language that everyone understands and can relate to; and 3. the “citability” and “findability” of publications – articles written in English and describing English are preferred over articles in/describing other languages and are therefore more cited, thus coming on top of lists of recommended articles.

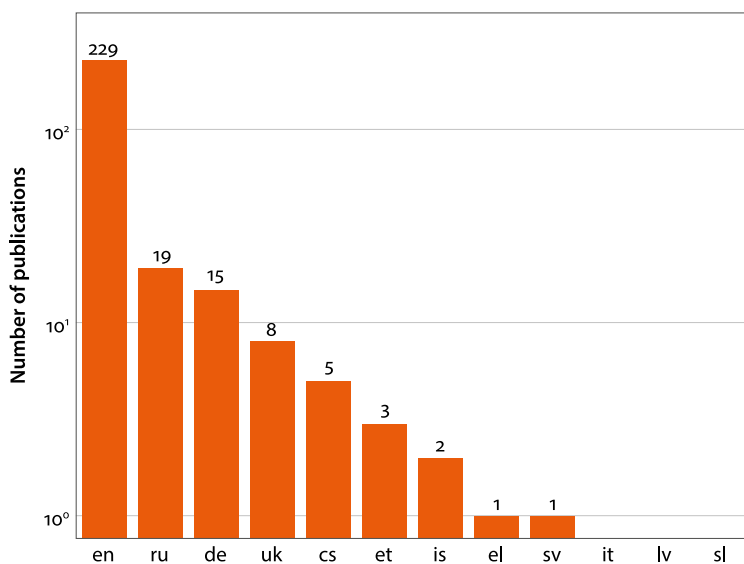


Figure 1. Number of GEC-related publications for the twelve different languages included in our dataset.² Figure from Masciolini et al. (2025b)

2. These estimates were obtained by searching the Association for Computational Linguistics (ACL) Anthology, looking for co-occurrences of the terms “grammatical error correction” or “GEC” with the names of our twelve languages in paper titles or abstracts. In cases where no MultiGEC language was mentioned explicitly, we manually filtered out papers dealing with languages outside of the scope of our dataset and reassigned papers that mentioned multilingualism or low-resource languages to the relevant language categories. Based on insights from Duce et al. (2022), all other papers leaving the language unspecified were assumed to deal with English. Conversely, we excluded and/or reassigned papers where English was explicitly mentioned, but clearly for reasons other than it being the main object of study, e.g. appearing in constructions such as “Beyond English GEC”. The search was performed on

Even though compounding, incremental advantages can significantly affect the opportunities and visibility of those affected negatively (i.e. “the poor”), affirmative actions can help close the gap. This has also been proposed in the recent “Second Language Acquisition (SLA) for All” initiative (Godfroid & Andringa, 2023), albeit for a different representational bias than ours. In the context of NLP, the dominance of English extends across several subfields, but making changes in general is a very broad and unrealistic goal, and efforts by individual researchers are less effective than internationally coordinated ones. This is why we launched the Computational SLA working group, an international cooperation with a special domain focus on Second Language Acquisition, and within which we work incrementally adopting a multilingual approach.³ Since its foundation in 2021, the group has supported and promoted work on less represented languages by building a community, releasing datasets and organizing shared tasks relevant for SLA and Learner Corpus Research (LCR) in the areas of, among others, Grammatical Error Detection (GED), Grammatical Error Correction (GEC) and essay grading.⁴

This paper introduces MultiGEC, which is a novel GEC dataset compiled for MultiGEC-2025.⁵ It represents the first text-level multilingual shared task on Grammatical Error Correction (Masciolini et al., 2025a); a cooperation between eight task organizers and 29 data providers.⁶ The goal of GEC is to improve texts by rewriting them in one of two possible ways: 1. reflecting the principle of ‘minimal correction’ and 2. applying fluency edits. Minimal corrections – traditional in Learner Corpus Research – are meant to produce texts that conform to the norms of the target language whilst preserving not only the intended meaning of the learner production, but also as much as possible of its grammar, lexis and writing style (e.g., Rudebeck & Sundberg, 2021). Fluency edits, on the other hand,

December 13, 2024. For an overview of the GEC history for the twelve MultiGEC languages, see Masciolini et al. (2025b).

3. spraakbanken.gu.se/en/compsla

4. In machine learning, a shared task is a competition aimed at tackling a specific research problem (the ‘task’). Typically, participants are provided with a dataset and invited to use it to train models to solve the task. Models are then comparatively evaluated, which in the end helps develop state-of-the-art approaches to solving the problem.

5. Given the computational nature of GEC and the fact that no metadata is included, we opted for referring to MultiGEC as a dataset rather than a corpus. Since it consists of a series of resources derived from pre-existing corpora, the terms ‘collection’ and ‘repository’ were also considered. However, we decided against them because of their inconsistent use in different NLP subfields. Finally, we discarded the term ‘benchmark’ as it is commonly used to refer to evaluation sets, whereas MultiGEC also includes training data.

6. spraakbanken.gu.se/en/compsla/multigec-2025

may also include more extensive rephrasing aimed at producing more idiomatic language (Sakaguchi et al., 2016). For instance, the text

in the pass when I have free time I go to see my laptop. In this, almost the music and movie.

could be minimally corrected to

In the past when I had free time I went to see my laptop. In this, there is almost all the music and movies.

whereas a fluency-edited version of the same text could be

I used to spend most of my free time on my laptop, where I had a lot of music and movies.

MultiGEC is the first highly multilingual dataset for text-level error correction. By bringing non-English datasets to the attention of the international NLP, SLA and LCR communities, we aim to foster an increased interest in low-resource languages. Moreover, the focus on full texts is meant to stimulate the development of GEC systems that are able to account for contexts larger than that of individual sentences, which has traditionally been the conventional unit of analysis for a variety of NLP tasks. While all MultiGEC subcorpora could already be obtained individually prior to the shared task, MultiGEC gathers the majority of them into a single convenient place.⁷ It also presents the dataset in a simple uniform plain-text format, which in turn enables cross-linguistically consistent processing and evaluation.⁸

2. Data

The MultiGEC dataset consists of seventeen subcorpora covering twelve different languages.⁹ Each subcorpus is a collection of texts, typically authored by learners of the target language, accompanied by one or more correction hypotheses.¹⁰ As mentioned in the Introduction, this dataset was created in the framework of a

7. doi.org/10.23695/h9f5-8143

8. With three exceptions when it comes to tokenization (cf. Table 1), all planned to be solved in a future version of the corpus.

9. doi.org/10.23695%2Fh9f5-8143

10. The entire corpus consists of full texts, with the exception of the Russian subcorpus RULEC-GEC, which consists of text fragments (cf. Section 2.9) and the Icelandic subcorpus IceL2EC, where some particularly long texts were split into shorter segments (cf. Section 2.6).

shared task and the correction hypotheses can thus be considered as the ‘gold-standard’ for the purpose of evaluating system hypotheses.

Table 1. Overview of the subcorpora that make up the MultiGEC dataset. This information is also available as machine-readable metadata

Language	Subcorpus	# essays	Hypothesis sets	Learners	Minimal	Fluency	Peculiarities
Czech	NatWebInf	6167	2	L1 (web)	✓		
Czech	Romani	3599	2	L1 (Romani children)	✓		
Czech	SecLearn	2407	2	L2	✓		
Czech	NatForm	391	2	L1 (students)	✓		
English	Write & Improve	5050	1	L2	✓		separate download
Estonian	EIC	258	3	L2	✓	✓	
Estonian	EKIL2	1503	2	L2		✓	
German	MERLIN	1033	1	L2	✓		pre-tokenized
Greek	GLCII	1289	1	L2	✓		
Icelandic	IceEC	176	1	L1 (mixed)		✓	pre-tokenized
Icelandic	IceL2EC	193	1	L2		✓	pre-tokenized; includes text fragments
Italian	MERLIN	813	1	L2	✓		
Latvian	LaVA	1015	1	L2	✓		
Russian	RULEC-GEC	6043	3	mixed (L2 + heritage)	✓	✓	pre-tokenized; includes text fragments; separate download
Slovenian	Solar-Eval	109	1	L1 (students)	✓		
Swedish	SweLL_gold	502	1	L2	✓		
Ukrainian	UA-GEC	1872	4	mixed (crowd-sourced)	✓	✓	

In Table 1, we give an overview of dataset statistics for MultiGEC, including the source corpora for each language, the number of texts and hypothesis sets (i.e., the number of distinct gold-standard correction versions) per corpus and various other characteristics.

As can be observed in the “learners” column, the subcorpora are in the majority of cases made up of essays written by L2 learners in an instructional setting. For some of the languages, however, the dataset also includes texts authored by L1 and heritage speakers, either collected together with the L2 productions or as a separate subcorpus. Given the low-resource nature of most of the languages, this author variation is somewhat inevitable. Despite our particular interest in SLA, however, this diversity is also valuable, as including different genres and author groups is indispensable for building a comprehensive GEC resource. Thanks to documentation and machine-readable metadata, system developers can easily select the subsets of MultiGEC that are best suited for their target users and their particular domain of application.

For roughly half of the subcorpora, at least some of the essays present multiple correction hypotheses, while the remaining half only comes with one per text (cf. column “hypothesis sets”). We were not able to identify corpora with multiple correction hypotheses for all languages. This is not ideal, but it is unsurprising since annotation is an extremely time-consuming and costly process. When possible, however, we opted for including all corrections since the reliability of many automatic GEC evaluation metrics increases when multiple hypotheses are available.

Another point of difference is correction style (columns “minimal” and “fluency”). While the majority of the subcorpora adopts a ‘minimal edits’ approach, three of them are fluency-edited and three more present correction hypotheses in both styles. Which approach is more suitable for GEC datasets is a matter of debate (cf. Sakaguchi et al., 2016): while minimal edits are grounded in pedagogy, the increasing use of Large Language Models (LLMs) in GEC research might be leading towards a shift towards fluency rewrites (Davis et al., 2024). Regardless of which strategy may prevail in the future, we see value in both minimal-editing GEC systems, which can be especially useful for learners facing the grammar of a new language, and in tools for general text improvement, targeting more advanced learners and/or the general population. For this reason, our shared task was organized into two separate tracks – one per correction style – leaving the choice of which subcorpora and hypotheses to use to participants, depending on their target users. In an ideal future, MultiGEC would grow to include correction hypotheses in both styles for each of the subcorpora. Note that correction hypotheses are the only kind of error annotation available as part of the MultiGEC dataset: error codes, even when present in the source subcorpora, are not included as they are neither necessary for GEC nor comparable across languages.

Furthermore, each subcorpus is split into a training, a development and a test set. For languages whose source corpora came pre-split, the MultiGEC version maintains the original partitions. When this was not the case, we opted for the commonplace 80:10:10 splitting strategy, as long as this allowed us to reserve a minimum of 40 essays for testing. Smaller subcorpora underwent a 10:45:45 subdivision. The rationale behind this is that we expect a majority of systems that use MultiGEC to be based on LLMs; a scenario in which few-shot learning could already prove effective. The reliability of evaluation (both at the testing stage and during development), on the other hand, greatly depends on the amount of available texts. The size of the different subcorpus splits, as well as more detailed statistics for each specific subcorpus, are presented on the dataset's dedicated webpage, as well as in the online supplementary material.¹¹

Each subcorpus split consists of a set of files aligned at the essay level: one file containing the original learner essays and one or more containing the corresponding corrected versions. Each file follows a simple Markdown-based format, designed to be both machine- and human-readable. Each essay is assigned an identifier to allow matching with the corresponding section of the relevant hypothesis file(s). No other metadata is included. An excerpt from the *English Write & Improve Corpus* (Nicholls et al., 2024) is displayed in Figure 2.

```
### essay_id = essay_26403d8120237fe2
```

In my free time. I playing video game, read a book and watch TV. The name of book is, "what is your parachute color," Thing I like doing playing video game on Friday. It is important to me. In this time I forget all things. I going to inside the game. The time go fast. On weekend I have a lot of time to do other things, but I can't do that thing.

```
### essay_id = essay_f80457440150ad1e
```

I can't deny that I am using a computer quiet often however the only thing I need on it is a browser. I prefer to use browser on a computer rather than on my phone.

The most enjoyable activities for me are listening to a music and sometimes watching something funny to steam off some pressure. I used to play a lot until I was fifteen. Back then something happened and I haven't played games ever since.

Figure 2. Excerpt from the training split of the *Write & Improve* subcorpus (original essays). Each text is preceded by a header with an essay identifier, through which it can be matched to its correction hypotheses

11. spraakbanken.github.io/multigec-2025

Whenever possible, we preserve the original spacing, with the addition of an empty line at the end of each text to visually separate subsequent essays. Deviations from this standard are listed under “peculiarities” in Table 1. Together with the data, we provide scripts to validate, parse and generate files in this format, as well as an LLM-based baseline and the automatic evaluation scripts that were used in the shared task.¹²

Access to the MultiGEC dataset is personal. Before downloading, new users are required to fill in a form through which they agree to the Terms of Use for the dataset.¹³ These define a few restrictions on its usage when it comes to re-identification of data subjects, use with proprietary API-based models, redistribution and commercialization. In addition, each MultiGEC subcorpus is subject to the license of the source corpus.¹⁴ Further details about the source corpora that were consulted for each language are provided in the following language-specific subsections.

2.1 Czech

Original essay	Correction hypothesis
Esej – Rady kamarádovi který se chce učít česky. Když chceš se učít česky musíš mít hodně čas, proto čeština je těžká. Každý den studuj minimalně jedna hodina, trenuj novou gramatiku, studuj novou slovu.	Esej – Rady kamarádovi, který se chce učít česky. Když se chceš učít česky, musíš mít hodně času, protože čeština je těžká. Každý den studuj minimálně jednu hodinu, trénuj novou gramatiku, studuj nová slova.
Approximate translation: “Essay – Advice for a Friend Who Wants to learn Czech. If you want to learn Czech, you’ll need plenty of time, because Czech is difficult. Study at least one hour every day, practice new grammar and learn new words.”	

Figure 3. Excerpt of an essay from the Czech *SecLearn* subcorpus. Corrected segments are highlighted in bold

The Czech portion of the MultiGEC dataset is directly derived from the *Grammar Error Correction Corpus for Czech* (GECCC) (Náplava et al., 2022).¹⁵ The full corpus consists of 12.6 thousand texts. Whereas GECCC is a single

12. github.com/spraakbanken/multigec-2025/tree/main/scripts

13. lt3.ugent.be/resources/multigec-dataset/dataset-download-form

14. For subcorpora derived from resource whose license terms are more permissive than the Terms of Use for MultiGEC, the former have precedence.

15. The GECCC corpus is available at hdl.handle.net/11234/1-4861

corpus consisting of sentence-level parallel files accompanied by metadata specifying, among others, the domain of each text, we decided to derive four domain-specific subcorpora for MultiGEC: *SecLearn*, *NatForm*, *Romani* and *NatWebInf*.

The *SecLearn* subcorpus includes essays written by non-native L2 learners of Czech, originally from the *CzeSL* (Rosen et al., 2020) and *MERLIN* (Boyd et al., 2014) corpora. The texts are representative of a wide variety of L1s and all proficiency levels in Czech, ranging from ‘hardly discernible from native’ to ‘nearly incomprehensible’.

The *NatForm* subcorpus includes transcripts of handwritten essays by native L1 Czech students of elementary and secondary schools from the *SKRIPT 2012* learner corpus (Šebesta et al., 2016). Depending on the age and competence of the student, the essays may be more or less close to standard Czech.

The *Romani* subcorpus includes the Romani ethnolect of Czech, consisting of transcripts of essays hand-written by schoolchildren with a Romani family background. The texts were taken from the *ROMi* corpus (Šebesta et al., 2014) and the *ROMi* section of the AKCES-GEC corpus.

The *NatWebInf* subcorpus consists of parts of informal discussions from Facebook and from the news server *Novinky.cz*, thus mostly representing adult native L1 speakers of Czech.¹⁶ Due to frequent misspellings and missing diacritics, however, these texts may still not be considered standard Czech.

Each text is paired with either one or two corrected versions. Annotators were instructed to correct grammar and spelling, with larger rephrasing only allowed in cases where the original text would otherwise appear incoherent. An example together with the corrected version and a translation is provided in Figure 3. All the source Czech corpora are available under a Creative Commons Attribution-Share Alike license (CC BY-SA 4.0).

2.2 English

The English portion of the MultiGEC dataset is derived from the *Write & Improve Corpus 2024* (Nicholls et al., 2024), published by Cambridge University Press & Assessment. Twenty-two different L1s feature in the corpus as a whole, including a typologically diverse range of languages such as Portuguese, Thai, Persian and Japanese. Essay prompts were selected so that 40% of essays were written in response to beginner level prompts; another 40% in response to intermediate level prompts and the remaining 20% in response to prompts devised for advanced learners. This distribution of difficulty levels was designed to broadly represent the user-base of *Write & Improve*, which mainly includes beginner and

16. novinky.cz

intermediate level learners of English. The *Write & Improve Corpus 2024* is available from the *English Language iTutoring website*,¹⁷ where applicants may read and agree to the license terms.

Within MultiGEC, the English dataset contains 5,050 essays written by learners of English using the *Write & Improve* platform,¹⁸ which gives automated grades and error feedback. The essays included in the MultiGEC dataset are the final versions of essay sets written by selected *Write & Improve* users in response to one of fifty selected prompts in the time period 2020–2022. The earlier versions of these essays are available, along with other annotations and metadata, as part of the larger *Write & Improve Corpus*. This subset of 5,050 essays has been corrected using a minimal edit approach; an example essay together with the minimal edits is shown in Figure 4.

The *Write & Improve* application is designed to encourage users to iteratively refine their text in order to improve it. Given the way that the platform tends to give feedback on commonly seen errors, relatively frequent errors such as missing articles or incorrect prepositions may be corrected by the learner by the time they submit the final version of their essay. The task for NLP systems is therefore to identify and correct the remaining, rarer or more complex kinds of errors.

The random assignment of essays to train, develop and test was made so that no single user features across splits. It also means that each one is representative of the overall mixture of L1s and proficiency levels in the main corpus.

Original essay	Correction hypothesis
Hello Robin! I am sorry to hear about you.	Hello Robin! I am sorry to hear about you.
How are you now? You got relief from your pain and how many weeks are you in the bed?	How are you now? You got relief from your pain, and how many weeks are you in bed? I
I wish you will well soon.	hope you will be well soon.

Figure 4. Excerpt of an essay from the English *Write & Improve Corpus 2024*. Corrected segments are highlighted in bold

2.3 Estonian

The Estonian MultiGEC dataset is derived from two error-annotated learner corpora: a subset of the *Estonian Interlanguage Corpus* (EIC) and the *L2 learner parallel error corpus of the Institute of Estonian Language* (EKIL2).

17. englishlanguageitutoring.com/datasets/write-and-improve-corpus-2024

18. writeandimprove.com

EIC is a growing corpus compiled at Tallinn University.¹⁹ It comprises texts written by learners of Estonian as an L2 in various educational settings, as well as L1 reference material. Its L2 subset, which was used as the basis for the corresponding MultiGEC subcorpus, consists of 258 L2 texts, also available in an error-annotated format.²⁰ It includes descriptive, narrative and argumentative writings, as well as both formal and informal letters representing proficiency levels ranging from A2 to C1. The material is licensed under the Creative Commons Attribution license (CC BY 4.0).

The *EKI Estonian L2 learner parallel error corpus* (EKIL2) consists of 1,503 learner essays.²¹ The dataset was automatically converted from the *EKI Error-annotated Estonian L2 learner corpus*, which contains data from 7th grade assessment tests (corresponding to level A2), 9th grade final exams (B1) and 12th grade state exams (B2). The material was manually pseudonymized and is licensed under the CLARIN ACA license.

For both corpora, annotators detected, error-coded and corrected various types of grammatical, orthographic and lexical errors. Parallel texts were automatically generated from this annotation and manually reviewed. An example essay together with the correction, as well as an English translation, is presented in Figure 5.

2.4 German

The German MultiGEC dataset is derived from the *MERLIN* corpus (Boyd et al., 2014; Wisniewski et al., 2013). *MERLIN* is an error-annotated written L2 learner corpus for German, Italian and Czech. The full multilingual corpus, created within the *MERLIN* project (2012–2014), is available at merlin-platform.eu and subject to the CC BY-SA 4.0 license. The texts in *MERLIN* were taken from standardized language tests and are methodologically related to the Common European Framework of Reference for Languages (Council of Europe, 2020).

All German learner essays are accompanied by a minimally corrected version, intended to fix orthographic and grammatical errors. An example, together with a correction and an English translation, is presented in Figure 6. The data splits used for the *MERLIN* corpus are taken from earlier work by Boyd (2018).

19. Corpus query is available at the Estonian Language Learning and Analysis Environment (ELLE): elle.tlu.ee

20. github.com/tlu-dt-nlp/EstGEC-L2-Corpus

21. metashare.ut.ee/repository/browse/eki-estonian-l2-learner-parallel-error-corpus/b4e4c37191f511ef85be77e08085ea557483da743a1243cba494f419df4dbf3a/

Original essay	Correction hypothesis
Tere, Liina. Kirjutan Sulle, sest mul on suur probleem. Minu külmkapp on katki ja vaja remontida. Ma oskan, et sinu abikaasa on väga hea meist- ter. Võib-olla ta saab aidata minule . Kui teil on vaba aeg, helista palun mulle telefonil – või tulge mulle külla aadressil Luha –, Tallinnas. Head aega.	Tere, Liina! Kirjutan Sulle, sest mul on suur probleem. Minu külmkapp on katki ja sega on vaja re- montida. Ma tean , et sinu abikaasa on väga hea meister. Võib-olla ta saab mind aidata . Kui teil on vaba aega, helistage palun mulle telefonil – või tulge mulle külla aadressil Luha –, Tallinn . Head aega!
Approximate translation: “Hello Liina, I am writing to you because I have a big problem. My refrigerator is broken and needs repair. I know that your husband is a good craftsman. Maybe he can help me. If you have spare time, please call me at – or come visit me at Luha street –, Tallinn. Goodbye! Your friend Olga”	

Figure 5. Example essay from the *Estonian EIC* subcorpus. Corrected segments are highlighted in bold

Original essay	Correction hypothesis
Ich möchte zu meinen Eltern fahren am Morgen . Deshalb Ich zu Haus werde nicht blieben bis Freitag . Kannst du mir helfen über mein Hund ? Ich brauche dass du mein Hund Essen geben , und spazieren mit es machen .	Ich möchte am Morgen zu meinen Eltern fahren . Deshalb werde ich bis Freitag nicht zu Haus bleiben . Kannst du mir mit meinem Hund helfen ? Ich brauche es, dass du meinem Hund Essen gibst , und mit ihm einen Spaziergang machst .
Approximate translation: “I want to go to my parents in the morning. Therefore I won’t stay at home till Friday. Can you help me with my dog? I need for you to feed my dog and to go on a walk with him.”	

Figure 6. Example essay from the German subcorpus based on *MERLIN*. Corrected segments are highlighted in bold. For the sake of compactness, the original spacing is not preserved here

2.5 Greek

The source corpus for the Greek MultiGEC dataset is the *Greek Learner Corpus II* (GLCII), a growing, partly error-annotated learner corpus designed to enhance research on Greek as a second (L2) or foreign language (FL) (Tantos et al., 2023). GLCII is currently the largest freely available corpus for L2 Greek, developed within the framework of the Latent Aspects in L2 Acquisition (LAL2A) project. GLCII includes 1,700 written and spoken productions comprising ~500,000-word tokens from L2 Greek learners. Learners contributing to GLCII

come from diverse sociolinguistic and cultural backgrounds and have attended Greek language courses in various settings within Greece and abroad. GLCII also includes a control corpus of L1 Greek, not included in MultiGEC. The GLCII dataset is freely accessible under a CC BY-SA 4.0 license.

All Greek learner essays have been minimally corrected addressing orthographic and grammatical deviations from native speaker production. An example, together with a correction and English translation is presented in Figure 7.

Original essay	Correction hypothesis
Χαίρουμε πολύ έπιηρες πτυχίο σου. Θα είναι που καλά αν ήθελες να κάνεις μεταπτυχιακό. Αλλά τώρα ακούσα εσύ θέλεις να δουλέψεις. Επιδή εσύ έναν δάσκαλος, να στείλεις το βιογραφικό σου στο κάποια το σχολεια.	Χαίρομαι πολύ που πήρες το πτυχίο σου. Θα ήταν καλό αν ήθελες να κάνεις μεταπτυχιακό. Αλλά τώρα άκουσα ότι θέλεις να δουλέψεις. Επειδή είσαι δάσκαλος, να στείλεις το βιογραφικό σου σε κάποια σχολεία.
Approximate translation: "I am glad you graduated. It would be good if you wanted to pursue a master's degree. But now I heard that you want to work. Since you are a teacher, send your resume to some schools."	

Figure 7. Example essay from the Greek subcorpus based on GLCII. Corrected segments are highlighted in bold. For the sake of compactness, the original spacing is not preserved here

2.6 Icelandic

The Icelandic contribution to the MultiGEC dataset is derived from two annotated error corpora: the *Icelandic Error Corpus* (IceEC) (Ingason et al., 2021) and the *Icelandic L2 Error Corpus* (IceL2EC) (Ingason et al., 2022).

IceEC focuses on native Icelandic texts, including high school student essays, online news texts and Wikipedia articles (Arnardóttir et al., 2021). The MultiGEC dataset only includes the student essays subset.

IceL2EC includes texts by adult Icelandic learners across CEFR levels, corrected by experienced instructors. The corpus is described in Glišić and Ingason (2022). The original essays include diverse written assignments of varying lengths, from 150–200-word beginner texts to several-thousand-word advanced essays (such as full MA theses), leading to an uneven word distribution. To create a more balanced dataset for model training, the longer texts were chunked into 40–50 sentence segments, resulting in a total of 193 texts with a more even word distribution.

These two corpora, both freely accessible under a CC BY-SA 4.0 license, are the first ever published Icelandic error corpora that include manually corrected texts annotated for spelling, grammar and fluency errors by professional linguists.

Both are available as layered TEI-format XML documents, pre-tokenized and error-coded, but have been converted back to plain text – only maintaining tokenization – to match the MultiGEC format.

The annotation process was meant to produce target hypotheses that align with grammatical norms and stylistic conventions of standard Icelandic, including substantial rephrasing wherever necessary. Consequently, both Icelandic subcorpora adhere to the ‘fluency edits’ principle. An example essay-correction pair is shown in Figure 8.

Original essay	Correction hypothesis
Bókin heitir er Blómin á Þakinu. Bókin gefin er út á nitján hundruð áttatíu og fimm . Bókin er teiknar þær er Brian Pilkington og sagan er eftir Ingibjörgu Sigurðardóttur.	Bókin heitir Blómin á þakinu. Bókin var gefin út 1985 . Bókin er teiknuð af Brian Pilkington og sagan er eftir Ingibjörgu Sig- urðardóttur.
Approximate translation: “The book is called The flowers on the roof. The book was published in 1985. The book is drawn by Brian Pilkington and the story is by Ingibjörg Sigurðardóttir.”	

Figure 8. Excerpt of an essay from the Icelandic subcorpus based on Icel2EC. Corrected segments are highlighted in bold. For the sake of compactness, the original spacing is not preserved here

2.7 Italian

Like the German subcorpus, the Italian portion of the MultiGEC dataset is derived from the *MERLIN* corpus (Boyd et al., 2014; Wisniewski et al., 2013). It contains 813 learner essays produced during official language tests, one of which displayed alongside its correction in Figure 9. For descriptions of the annotations and metadata provided in *MERLIN*, as well as for the license conditions for the source corpus, please refer to Section 2.4.

Original essay	Correction hypothesis
Ciao Petter Tutto bene? Io vogli aiutarti. Dove lavori adesso? Il tuo lavoro non ti piace piu?	Ciao Petter, tutto bene? Io voglio aiutarti. Dove lavori adesso? Il tuo lavoro non ti piace più?
Adeso sei cenca lavoro? Devi iniziare subito?	Adesso sei senza lavoro? Devi iniziare subito?
Cosa fai fare? Aspeto la tua respasta. Michele	Cosa sai fare? Aspetto la tua risposta. Michele
Approximate translation: “Hello Petter, everything alright? I want to help you. Where do you work now? Do you not like your work anymore? Are you without work now? Do you have to start immediately? What work can you do? I am waiting for your reply. Michele”	

Figure 9. Example essay from the Italian subcorpus based on *MERLIN*. Corrected segments are highlighted in bold

2.8 Latvian

The Latvian subset of MultiGEC comprises 1015 texts and correction hypotheses (see Figure 16) from the *Latvian Language Learner corpus* (LaVA) of learner-written essays (Dargis et al., 2020, 2022). The source corpus features texts authored by learners with several different mother tongues studying at Latvian higher education institutions for the first or second semester, reaching beginner levels (approx. A1 and A2 at the CEFR scale) and is released under a CC BY 4.0 license.

In terms of general preprocessing, all essays were morphologically annotated and personal information was removed. Learner errors were then manually corrected and annotated using minimal correction edits. An example essay-correction pair, together with an English translation, is presented in Figure 10.

Original essay	Correction hypothesis
Mani sauc Fran, es esmu no Vācijā. Mana ģimene ir divdesmit cilvēki. Man ir liela ģimene. Man garšo pica. Man negaršo Kiwi. Man ir trīsdesmit septiņi. Es studēju medicīna. Man patīk šeit. Es dzīvoju kopā ar trīs draugiem.	Mani sauc Frena, es esmu no Vācijas. Manā ģimenē ir divdesmit cilvēki. Man ir liela ģimene. Man garšo pica. Man negaršo kivi. Man ir trīsdesmit septiņi gadi . Es studēju medicīnu. Man patīk šeit. Es dzīvoju kopā ar trīs draugiem.
Approximate translation: “My name is Frena, I’m from Germany. My family consists of twenty people. I have a big family. I like pizza. I don’t like kiwi. I am thirty-seven years old. I am studying medicine. I like it here. I live with three friends.”	

Figure 10. Example essay from the Latvian subcorpus based on LaVA. Corrected segments are highlighted in bold

2.9 Russian

The Russian subcorpus of the MultiGEC dataset is derived from RULEC-GEC (Rozovskaya & Roth, 2019). RULEC-GEC is an error-annotated corpus based on a subset of the *Russian Learner Corpus of Academic Writing* (RULEC, Alsufieva et al., 2012). 77% of the RULEC-GEC corpus consists of essays and papers written by 15 students learning Russian as a foreign language in a university setting in the United States. The remaining part is composed of texts authored by 13 heritage speakers who grew up in the United States but had exposure to Russian at home. The corpus is openly available through a CC BY SA 4.0 license. The original RULEC-GEC train/dev/test splits are retained in MultiGEC.

While the original version of RULEC-GEC contains one correction hypothesis per sentence, its test split has been recently enriched with two additional cor-

rections per sentence, for a total of three corrections per sentence (Palma Gomez & Rozovskaya, 2024). Both annotation paradigms follow the minimal-edit principle. Figure 11 illustrates a sample snippet and its approximate English translation.

As the source corpus is annotated at the sentence, rather than essay level, integration into the MultiGEC dataset required reconstructing text snippets from consecutive RULEC-GEC sentences. Additionally, it must be noted that the data comes pre-tokenized as a result of the format in which the source corpus is stored.

Original essay	Correction hypothesis
<p>Дорогие геронтологи, Мы приглашаем вас в конференции в Турции называется “Экономические развития городов в XXIом веке”. Мы лично знаем какой важные роль пожилые люди играют в мировом сообществе, и так конференция включает круглый стол о “науке старения”.</p>	<p>Дорогие геронтологи, Мы приглашаем вас на конференцию в Турции, которая называется “Экономическое развитие городов в XXIом веке”. Мы знаем, какую важную роль пожилые люди играют в мировом сообществе, конференция также включает круглый стол о “науке старения”.</p>
<p>Approximate translation: “Dear gerontologists, We invite you to a conference in Turkey called “Economic development of cities in the 21st century”. We personally know the important role older people play in the global community and so the conference includes a roundtable on the “science of aging”.”</p>	

Figure 11. Excerpt of a text snippet from the Russian subcorpus based on RULEC-GEC. For the sake of readability, the present text was manually untokenized

2.10 Slovene

The Slovene portion of the MultiGEC dataset is derived from the *Šolar-Eval 1.0 corpus* (Gantar et al., 2023), an evaluation dataset created for assessing automated spelling and grammar-correction systems in standard Slovene. *Šolar-Eval 1.0* originates from the broader *Šolar 3.0 corpus* (Arhar Holdt & Kosem, 2024), which contains 5,485 essays written by native Slovene speakers – primary and secondary school students – along with 36,570 authentic teacher corrections as they occurred in the educational setting. While these teacher corrections provide valuable insights into feedback patterns within the context of student language development, inconsistencies in their application limited the suitability of the corpus for the NLP domain.

Šolar-Eval was developed to address this limitation, offering a curated subset of 109 essays with consistent corrections carefully applied according to the principle of minimal correction (Arhar Holdt et al., 2023). An example of a corrected text is provided in Figure 12. The dataset is openly accessible through a CC BY

NC-SA 4.0 license. For inclusion in MultiGEC, the dataset without the above-mentioned metadata is split into training, validation and test sets.

Original essay	Correction hypothesis
Ta zgodba pripoveduje o Odiseju in njegovih mož se srečajo z kiklopom, ki jih zadržuje v njegovi jami. Zato si je Odisej izmislil pameten in prebrisan načrt, da bo oslepil Polifema . To mu je uspelo tako, da je ošpičil olkovo deblo in jo dal v žerjavico.	Zgodba pripoveduje o Odiseju in njegovih možeh. Srečajo se s kiklopom, ki jih zadržuje v svoji jami. Zato si je Odisej izmis- lil pameten in prebrisan načrt, da bo Polifema oslepil . To mu je uspelo tako, da je ošpičil oljkovo deblo in ga dal v žerjavico.
Approximate translation: “The story tells of Odysseus and his men. They meet a Cyclops who keeps them trapped in his cave. Therefore, Odysseus devised a smart and cunning plan to blind Polyphemus. He succeeded by sharpening an olive trunk and placing it in the embers.”	

Figure 12. Example from the Slovene subcorpus based on Šolar-Eval. Corrected segments are highlighted in bold

2.11 Swedish

The Swedish portion of the MultiGEC dataset is derived from the *Swedish Learner Language corpus*, SweLL-gold (Volodina et al., 2019, 2022). SweLL-gold consists of 502 essays written by adult learners of L2 Swedish, a group presenting great diversity in terms of language and educational background, as well as when it comes to proficiency level in the target language. The source corpus comes with rich metadata and annotation: each essay is manually pseudonymized, corrected and error-tagged. In addition, both the original and the corrected versions of each text are linguistically annotated with the Sparv pipeline (Hammarstedt et al., 2022). SweLL-gold is distributed through Språkbanken (the national Language Bank of Sweden), and subject to the CLARIN-ID -PRIV -NORED -BY license.^{22, 23}

The stated objective of the error annotation process is to produce corrected versions of the essays which follow the norms of standard Swedish whilst trying to preserve the intended meaning and writing style of the original texts, with more extensive rephrasing only allowed when necessary from a grammatical acceptability standpoint. For this reason, SweLL-gold correction hypotheses are to be considered as minimal edits. An example essay-correction pair accompanied by a translation is given in Figure 13.

In the MultiGEC version of the corpus, full texts preserving original spacing were reconstructed from SweLL-gold XML files, discarding metadata and all of the above-mentioned annotation layers.

22. doi.org/10.23695/2k47-y432

23. kielipankki.fi/support/clarin-eula

Original essay	Correction hypothesis
bor i Tuna, ser ut normal tycker jag min mamma hennes man och min lilla bror alla vi bor tillsammans och vi bor typ 10 min nära tunnelbana när jag bodde i mitt land de va nasta samma men jag bodde med min murmmor bara	Jag bor i Tuna, det ser normalt ut tycker jag. Min mamma, hennes man och min lilla bror, alla vi bor tillsammans och vi bor typ 10 min från tunnelbanan. När jag bodde i mitt land var det nästan likadant men jag bodde med min mormor bara.
Approximate translation: "I live in Tuna, it looks normal I think. My mum, her husband and my little brother, we all live together and we live like 10 min away from the subway. When I lived in my country it was almost the same but I lived with my grandma only."	

Figure 13. Example essay from the Swedish subcorpus based on SweLL-gold. Corrected segments are highlighted in bold. For the sake of compactness, the original spacing is not preserved here

2.12 Ukrainian

The Ukrainian portion of the MultiGEC dataset is derived from UA-GEC, an error-annotated corpus for Ukrainian GEC. UA-GEC encompasses texts from 828 contributors representing diverse backgrounds (Syvokon et al., 2023). The contributors donated their texts, including social media posts, essays, technical documents, advertisements, chat messages, translations and more. The source corpus maintains anonymized metadata about contributors, including their region, occupation, gender and native speaker status. UA-GEC is distributed under the CC-BY 4.0 license, making it freely available for research and commercial use.

The annotation process was conducted by professional linguists, all native Ukrainian speakers holding degrees in Ukrainian linguistics. All texts in the development and test sets have two correction hypotheses, whereas training instances are paired with a single correction. Corrections originally included both grammatical and fluency edits. Each edit was manually classified into 22 error type categories. To create minimally edited corrections, fluency edits were reverted and the resulting texts were manually proofread to ensure that the process does not leave any sentence in a broken state. An example is presented in Figure 14.

The original UA-GEC dataset is split into a training and a test set. The UNLP 2023 Shared Task (Syvokon & Romanushyn, 2023) further split the original test set into smaller development and test sets. Contributions from a single author are assigned exclusively to one split. The MultiGEC dataset maintains the same train-dev-test structure.

Original essay	Correction hypothesis
<p>Нещодавно, завдяки Filipa Carvalho Marques, мені до рук потрапила дивовжна стаття. Стаття представляє результати моделювання розповсюдження коронавірусу у Великій Британії та у США для різних сценаріїв протидії.</p>	<p>Нещодавно завдяки Filipa Carvalho Marques мені до рук потрапила дивовжна стаття. Стаття представляє результати моделювання поширення коронавірусу у Великій Британії та у США для різних сценаріїв протидії.</p>
<p>Approximate translation: “Recently, thanks to Filipa Carvalho Marques I got my hands on an amazing article. The article presents the simulation results of the spread of the coronavirus in Great Britain and the USA for different countermeasure scenarios.”</p>	

Figure 14. An essay excerpt from the Ukrainian subcorpus derived from UA-GEC

3. Conclusions and future outlook

We have introduced MultiGEC, the first multilingual GEC dataset for text-level error correction, featuring no less than twelve languages. The MultiGEC-2025 shared task serves as a launch pad for the dataset, but the latter remains available after the end of the competition itself.

While the differences between MultiGEC subcorpora in terms of size, authorship and correction style pose a limit to the extent to which results can be compared across its twelve languages, this diversity is also one of the main strengths of the dataset. By providing easy access to data from a heterogeneous collection of GEC corpora in a simple and uniform format, we allow researchers to focus on the technical challenges of error correction rather than data sourcing and wrangling. Through machine-readable metadata and extensive documentation, we allow system developers to select the subcorpora that are best suited for their domain of application and user base.

In the future, we hope to see MultiGEC grow to encompass even more languages and text sources, as well as to become increasingly balanced in terms of subcorpora sizes and correction styles. The goal is for the dataset to turn into an increasingly more comprehensive and high-quality resource to support NLP researchers in the development of GEC systems targeting different user groups and domains of application for a wide variety of languages.

The MultiGEC dataset and the 2025 shared task are part of the Computational SLA working group’s ongoing effort to counteract the *Matthew effect* in NLP. Future work will broaden the focus to other subfields and tasks, such as automatic essay grading and CEFR level classification in multiple languages.

Funding

Open Access publication of this article was funded through a Transformative Agreement with University of Gothenburg.

Czech Compilation of the Czech part from existing resources was supported by the Czech Ministry of Education, Youth and Sports (grant no. LM2023044: Large Research, Development and Innovation Infrastructures). The task would be much harder without the source GECCC corpus, due to Jakub Náplava, Milan Straka and Jana Straková. Moreover, the result would be much smaller, less varied, or non-existent without previously built learner corpora, due to Karel Šebesta, Svatava Škodová, Barbora Štindlová, Jirka Hana and many others.

English Andrew Caines has been supported by Cambridge University Press & Assessment. Thanks to Diane Nicholls and Paula Buttery for co-preparation of the *Write & Improve Corpus* 2024.

Estonian We thank the following people for their contribution in compiling and annotating the corpus material: Pille Eslon, Kaisa Norak, Karina Kert, Silvia Maine and Linda Luig for their work on the EIC subcorpus; Kadri Sõrmus, Jelena Kallas, Sven Aller, Helen Kaljumäe, Silver Vapper, Anita Väli, Karoliina Jõgi and Marta Kohv for their work on the EKIL2 corpus and its source corpus EMMA; Krista Liin for consulting the work on both datasets. The work on EKIL2 dataset is co-funded by the European Union.

German and Italian The MERLIN project was funded from 2012 until 2014 by the EU Lifelong Learning Programme under project number 518989-LLP-1-2011-1-DE-KA2-KA2MP.

Greek The compilation of the Greek Learner Corpus II has been supported by the Hellenic Foundation for Research and Innovation through the project “Latent Aspects in L2 Acquisition (LAL2A)” (Grant Number 3161) as part of the 1st Call for “Research Projects to Support Faculty Members & Researchers and Procure High-Value Research Equipment”.

Icelandic The error corpora project was funded by the Icelandic Government as a part of the Language Technology Programme for Icelandic 2019–2023. We thank the following people for their contribution in collecting, correcting, and annotating the corpora: Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, Dagbjört Guðmundsdóttir, Isidora Glišić and Xindan Xu.

Latvian Work on Latvian has been supported by the State Research Programme’s project LATE (grant agreement No. VPP-LETONIKA-2021/1-0006), which is in synergy with the Latvian Council of Science grant Common Writing Errors in Latvian (lzp-2023/1-0481).

Slovene The research program Language Resources and Technologies for Slovene (P6-0411) and the projects Empirical Foundations for Digitally-Supported Development of Writing Skills (J7-3159) and Large Language Models for Digital Humanities (GC-0002) are funded by the Slovenian Research and Innovation Agency.

Swedish Work on Swedish has been supported by Nationella Språkbanken and Huminfra, both funded by the Swedish Research Council (2018–2024, contract 2017-00626; 2022–2024, contract 2021-00176) and their participating partner institutions, as well as the Swedish Research Council grant 2019-04129.

Ukrainian The creation of UA-GEC was initiated and supported by Grammarly. We extend special gratitude to Olena Nahorna, Pavlo Kuchmiichuk, Nastasiia Osidach, Ira Kotkalova, Anna Vesnii, Halyna Kolodkevych, and everyone else who participated in the corpus creation.





Open data badge and data availability statement






This article has been awarded an open data badge. MultiGEC data (DOI: doi.org/10.23695/Fh9f5-8143) can be downloaded via Ghent University at lt3.ugent.be/resources/multigec-dataset upon agreeing to its Terms of Use, with the following exceptions:










- human-written correction hypotheses for the test sets of all subcorpora, which are kept private to allow fair evaluation of future GEC models;
- the English Write & Improve subcorpus, which can be downloaded separately from the ELiT website (englishlanguageitutoring.com/datasets/write-and-improve-corpus-2024);
- the Russian RULEC-GEC subcorpus, which can be obtained by contacting Alla Rozovskaya (sigaliyah@gmail.com).

Learn more about the Open Practices badges from the Center for Open Science: <https://osf.io/tvyxz/wiki>

References


-  Alsufieva, A., Kisselev, O., & Freels, S. (2012). Results 2012: Using flagship data to develop a Russian learner corpus of academic writing. *Russian Language Journal*, 62, 79–105.
- Arhar Holdt, Š., Gantar, P., Bon, M., Gapsa, M., Lavrič, P., & Klemen, M. (2023). Dataset for evaluation of Slovene spell- and grammar-checking tools Šolar-Eval 1.0. (Slovenian language resource repository CLARIN.SI). <https://www.cjvt.si/prop/en/>
-  Arhar Holdt, Š., & Kosem, I. (2024). Šolar, the developmental corpus of Slovene. *Language Resources and Evaluation*, 1–27.
- Arnardóttir, Þ., Xu, X., Guðmundsdóttir, D., Stefánsdóttir, L., & Ingason, A. (2021). Creating an Error Corpus: Annotation and Applicability. In *Proceedings of CLARIN 2021 Annual Conference* (pp. 59–63).
-  Bol, T., de Vaan, M., & van de Rijt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, 115(19), 4887–4890.
-  Boyd, A. (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* (pp. 79–84). Association for Computational Linguistics.
- Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., & Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1281–1288). European Language Resources Association (ELRA).
- Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors. *Council of Europe Publishing*.
- Darģis, R., Auziņa, I., Kaija, I., Levāne-Petrova, K., & Pokratniece, K. (2022). LaVA–Latvian Language Learner corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 727–731).

- Dargis, R., Auziņa, I., Levāne-Petrova, K., & Kaija, I. (2020). Quality focused approach to a learner corpus development. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 392–396).
-  Davis, C., Caines, A., Andersen, Ø., Taslimipour, S., Yannakoudakis, H., Yuan, Z., Bryant, C., Rei, M. & Buttery, P. (2024). Prompting open-source and commercial language models for grammatical error correction of English learner text. In *Findings of the association for computational linguistics: ACL 2024* (pp. 11952–11967). Association for Computational Linguistics.
- Ducel, F., Fort, K., Lejeune, G., & Lepage, Y. (2022). Do we name the languages we study? the #BenderRule in LREC and ACL articles. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odiijk & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 564–573). European Language Resources Association.
-  Gantar, P., Bon, M., Gapsa, M., & Arhar Holdt, Š. (2023). Šolar-Eval: Evalvacijska množica za strojno popravljanje jezikovnih napak v slovenskih besedilih. *Jezik in Slovtvo*, 68(4), 89–108.
- Glišić, I., & Ingason, A. K. (2022). The Nature of Icelandic as a second language: An insight from the Learner Error Corpus for Icelandic. In *Proceedings of the CLARIN Annual Conference* (p. 23–33).
-  Godfroid, A., & Andringa, S. (2023). Uncovering sampling biases, advancing inclusivity, and rethinking theoretical accounts in Second Language Acquisition: Introduction to the special issue SLA for all? *Language Learning*, 73(4), 981–1002.
- Hammarstedt, M., Schumacher, A., Borin, L., & Forsberg, M. (2022). Sparv 5 user manual (Tech. Rep.). *Språkbanken Text*.
- Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., & Xu, X. (2021). Icelandic Error Corpus (IceEC) Version 1.1. (CLARIN-IS).
- Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., Xu, X., Glišić, I., & Guðmundsdóttir, D. (2022). The Icelandic L2 Error Corpus (IceL2EC) 1.3 (22.10). (CLARIN-IS).
- Masciolini, A., Caines, A., De Clercq, O., Kruijsbergen, J., Kurfalı, M., Muñoz Sánchez, R., Volodina, E., Östling, R. (2025a). The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL. In R. Muñoz Sánchez, D. Alfter, J. Kallas, & E. Volodina (Eds.), *Proceedings of the 14th workshop on Natural Language Processing for Computer Assisted Language Learning*. Tallin, Estonia: University of Tartu. <https://hdl.handle.net/2077/84800>
- Masciolini, A., Caines, A., De Clercq, O., Kruijsbergen, J., Kurfalı, M., Muñoz Sánchez, R., ... Zesch, T. (2025b). An overview of grammatical error correction for the twelve MultiGEC-2025 languages. *GU-ISS Forskningsrapporter från Institutionen för svenska språket*. Institution for Swedish, Multilingualism, Language Technology; University of Gothenburg. <https://hdl.handle.net/2077/84800>
-  Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810), 56–63.
-  Náplava, J., Straka, M., Straková, J., & Rosen, A. (2022). Czech grammar error correction with a large and diverse corpus. *Transactions of the Association for Computational Linguistics*, 10, 452–467.

- Nicholls, D., Caines, A., & Buttery, P. (2024). *The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English*. Cambridge University Press Assessment.
- Palma Gomez, F., & Rozovskaya, A. (2024). Multi-reference benchmarks for Russian grammatical error correction. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (volume 1: Long papers) (pp. 1253–1270). Association for Computational Linguistics.
-  Perc, M. (2014). The Matthew effect in empirical data. *Journal of The Royal Society Interface*, 11(98), 20140378.
- Rosen, A., Hana, J., Hladká, B., Jelínek, T., Škodová, S., & Štindlová, B. (2020). *Compiling and annotating a learner corpus for a morphologically rich language – CzeSL, a corpus of non-native Czech*. Karolinum, Charles University Press.
-  Rozovskaya, A., & Roth, D. (2019). Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7, 1–17.
- Rudebeck, L., & Sundberg, G. (2021). SweLL correction annotation guidelines. (Tech. Rep.). GU-ISS Research report series, Department of Swedish, University of Gothenburg.
-  Sakaguchi, K., Napoles, C., Post, M., & Tetreault, J. (2016). Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4, 169–182.
- Šebesta, K., Bedřichová, Z., Šormová, K., Straňák, P., & Peterek, N. (2014). ROMi 1.0. (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University).
- Šebesta, K., Goláňová, H., Letafková, J., & Jelínková, B. (2016). AKCES 1. (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University).
-  Sogaard, A. (2022). Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 5254–5260). Association for Computational Linguistics.
-  Syvokon, O., Nahorna, O., Kuchmiichuk, P., & Osidach, N. (2023). UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian Language. In *Proceedings of the second Ukrainian Natural Language Processing workshop (UNLP)* (pp. 96–102). Association for Computational Linguistics.
-  Syvokon, O., & Romanyshyn, M. (2023). The UNLP 2023 Shared Task on Grammatical Error Correction for Ukrainian. In *Proceedings of the second Ukrainian Natural Language Processing workshop (UNLP)* (pp. 132–137). Association for Computational Linguistics.
-  Tantos, A., Amvrazis, N., & Drakonaki, E. (2023). Greek Learner Corpus II (GLCII): Design and development of an online corpus for L2 Greek. *Journal of Applied Linguistics*, 36, 125–150.
-  Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., ... & Wirén, M. (2019). The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology (NEJLT)*, 6, 67–104.
-  Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., ... & Wirén, M. (2022). SweLL-gold. *Språkbanken Text*. Distributed via SBX/CLARIN.

Wisniewski, K., Schöne, K., Nicolas, L., Vettori, C., Boyd, A., Meurers, D., ... Hana, J. (2013). MERLIN: An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data. In *International Conference, ICT for Language Learning*, 6th edition.

Address for correspondence

Arianna Masciolini
Språkbanken Text
Institution for Swedish, Multilingualism, Language Technology
University of Gothenburg
Box 200
40530 Gothenburg
Sweden
arianna.masciolini@gu.se
 <https://orcid.org/0009-0009-2008-5842>

Co-author information

- Andrew Caines
Department of Computer Science &
Technology
University of Cambridge
apc38@cam.ac.uk
- Orphée De Clercq
LT₃ Language and Translation Technology
Team
Ghent University
orphee.declercq@ugent.be
- Joni Kruijsbergen
LT₃ Language and Translation Technology
Team
Ghent University
joni.kruijsbergen@ugent.be
- Murathan Kurfali
RISE Research Institutes of Sweden AB
murathan.kurfali@su.se
- Ricardo Muñoz Sánchez
Språkbanken Text
Institution for Swedish, Multilingualism,
Language Technology
University of Gothenburg
ricardo.munoz.sanchez@gu.se
- Elena Volodina
Språkbanken Text
Institution for Swedish, Multilingualism,
Language Technology
University of Gothenburg
elena.volodina@gu.se
- Robert Östling
Institutionen för lingvistik
Stockholms Universitet
robert@ling.su.se
- Kais Allkivi
Tallinn University
kais@tlu.ee
- Špela Arhar Holdt
Faculty of Computer and Information
Science
University of Ljubljana
arharhs@ff.uni-lj.si
- Ilze Auzina
Institute of Mathematics and Computer
Science
University of Latvia
ilze.auzina@lumii.lv
- Roberts Dargis
Institute of Mathematics and Computer
Science
University of Latvia
roberts.dargis@lumii.lv
- Elena Drakonaki
Department of Linguistics
School of Philology
Aristotle University of Thessaloniki
chrysiel@lit.auth.gr
- Jennifer-Carmen Frey
Institute for Applied Linguistics
Eurac Research
jennifercarmen.frey@eurac.edu
- Isidora Glišić
Icelandic and Comparative cultural studies
University of Iceland
isidora@hi.is
- Pinelopi Kikilintza
Department of Linguistics
School of Philology
Aristotle University of Thessaloniki
pkikilin@lit.auth.gr
- Lionel Nicolas
Institute for Applied Linguistics
Eurac Research
lionel.nicolas@eurac.edu

Mariana Romanyshyn
Grammarly
mariana.romanyshyn@grammarly.com

Alexandr Rosen
Institute of Linguistics
Faculty of Arts
Charles University
alexandr.rosen@ff.cuni.cz

Alla Rozovskaya
Department of Computer Science
Queens College
City University of New York
arozovskaya@qc.cuny.edu

Kristjan Suluste
Institute of the Estonian Language
kristjan.suluste@eki.ee

Oleksiy Syvokon
Microsoft
osyvokon@microsoft.com

Alexandros Tantos
Department of Linguistics
School of Philology
Aristotle University of Thessaloniki
alextantos@lit.auth.gr

Despoina-Ourania Touriki
Department of Linguistics
School of Philology
Aristotle University of Thessaloniki
dtouriki@lit.auth.gr

Konstantinos Tsiotskas
Department of Linguistics
School of Philology
Aristotle University of Thessaloniki
ktsiotsk@lit.auth.gr

Eleni Tsourilla
Department of Linguistics
School of Philology
Aristotle University of Thessaloniki
tsourilla@lit.auth.gr

Vassilis Varsamopoulos
Department of Linguistics
School of Philology
Aristotle University of Thessaloniki
varsamopo@gmail.com

Katrin Wisniewski
Herder Institute for German as a Foreign/
Second Language
Leipzig University
katrin.wisniewski@uni-leipzig.de

Aleš Žagar
Faculty of Computer and Information
Science
University of Ljubljana
ales.zagar@fri.uni-lj.si

Torsten Zesch
CATALPA
FernUniversität in Hagen
torsten.zesch@fernuni-hagen.de

Publication history

Date received: 13 November 2024

Date accepted: 11 February 2025

Published online: 1 April 2025