Beyond Error Categories: A Contextual Approach of Evaluating Emerging Spell and Grammar Checkers

Þórunn Arnardóttir[1,2], Svanhvít Lilja Ingólfsdóttir[2], Haukur Barri Símonarson[2], Hafsteinn Einarsson[1], Anton Karl Ingason[1], Vilhjálmur Þorsteinsson[2]
[1]University of Iceland, [2]Miðeind ehf.
[1]Sæmundargata 2, 102 Reykjavík, Iceland, [2]Fiskislóð 31 B/303, 101 Reykjavík, Iceland
[1]{thar, hafsteinne, antoni}@hi.is
[2]{svanhvit, haukur, vt}@mideind.is

Abstract

Automatic spell and grammar checking can be done using various system architectures, and large language models have recently been used to solve the task with promising results. Here we describe a new method of creating test data to measure the performance of spell and grammar checkers, including large language models. Three types of test data represent different approaches to evaluation, from basic error detection to error correction with natural language explanations of the corrections made and error severity scores, which is the main novelty of this approach. These additions are especially useful when evaluating large language models. We present a spell and grammar checking test set for Icelandic in which the described approach is applied. The data consists of whole texts instead of discrete sentences, which facilitates evaluating context awareness of models. The resulting test set can be used to compare different spell and grammar checkers and is published under permissive licenses.

Keywords: test data, evaluation, spell and grammar checking, large language models, Icelandic

1. Introduction

Automatic spell and grammar checking deals with various spelling and grammar errors in text, typos, deviations from the accepted language standard, and stylistic flaws. Work on Icelandic spell and grammar checkers has evolved quickly in the last years (see Óladóttir et al. (2022)), but Icelandic is still considered low-resourced in the European language technology field (Rehm and Way, 2023), and test sets for Icelandic spell and grammar checkers are scarce. Methods for evaluating spell and grammar checking systems range from feedback from language experts to a fully automated approach based on a particular metric and test set (Napoles et al., 2016; Fang et al., 2023; Wu et al., 2023). Expert feedback can be hard to come by, so automatic evaluation methods are valuable tools.

Until now, evaluation data for spell and grammar checkers has been limited to sentences, corrected and annotated with predetermined error categories. However, the paradigm shift that emerges with the abilities of large language models (LLMs) opens up many options for creating better and more flexible spell and grammar checkers, calling for a re-examination of how evaluation data is prepared and applied.

Here we present a new method of creating test data for evaluating spell and grammar checkers, including modern LLM-based ones, both existing and emerging. The dataset consists of complete texts, which are manually annotated, and is in three parts, each one annotated differently, to better encompass strengths and weaknesses of the models evaluated, from simply detecting errors to explaining the corrections made. In particular, we present data where language experts correct errors in texts and annotate them with explanations as to why they make a particular change, using free-form text. In addition to explanations, severity scores are assigned to corrected errors. This is an effort to move away from typical test data, and towards more user-oriented data. Moreover, the demand for explainable AI has been increasing, and the method described here is a step towards better evaluation of such systems as they emerge. The test set is published under a permissive license (Símonarson et al., 2023).

2. Related Work

Within automatic spell and grammar checking, rule-based methods are being replaced by neural network-based methods. Solving the spell and grammar checking task as a machine translation task is a prevalent method (Yuan and Briscoe, 2016; Ji et al., 2017; Junczys-Dowmunt et al., 2018; Korre and Pavlopoulos, 2022). LLMs can be used for spell and grammar checking and models such as GPTs (Floridi and Chiriatti, 2020) and LLaMa (Touvron et al., 2023) have broader abilities than smaller models. They tend to be better at evaluating and correcting text fluency,

SIGUL2024 Workshop, pages 45–52
21-22 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

and they are in general good at finding errors in text, including context-dependent errors (Penteado and Perez, 2023; Li et al., 2023; Qu and Wu, 2023). However, they sometimes overcorrect text, paraphrasing it unnecessarily and detecting errors where there are none, which is not as common with state-of-the-art (SOTA) methods.

The spell and grammar checking task is largely language-dependent, and the most prominent and accessible spell and grammar checkers for Icelandic are a rule-based one (Óladóttir et al., 2022) and a byte-level neural network-based model (Ingólfsdóttir et al., 2023). While the rule-based method can detect syntactic inconsistencies and errors, and justify its discoveries, the byte-level model is more robust, capable of correcting texts with multiple and complex errors, but lacks explainability. LLMs capable of checking spelling and grammar are currently not available for Icelandic.

Recently developed test sets for evaluating spell and grammar checkers contain corrected texts, where errors have been annotated, either manually or automatically, corrected and often categorized into error types (see e.g. Wang et al. (2022); Bexte et al. (2022); Katinskaia et al. (2022) and Korre and Pavlopoulos (2022)). Some Icelandic error corpora have been published in recent years, with manually annotated errors which have been corrected and categorized by error type (Arnardóttir et al., 2021, 2022; Ingason et al., 2021b, 2022b,a). Commonly used automated evaluation metrics for spell and grammar error checkers include $F_{0.5}$ and GLEU (Wang et al., 2020). $F_{0.5}$ is based on the precision and recall metric but precision is given twice the weight of recall. This means that correctly corrected errors are prioritized over all possible errors being corrected. $F_{0.5}$ is included in ERRANT (Bryant et al., 2017) and was used in the CoNLL-2014 shared task (Ng et al., 2014). The GLEU score rewards correct edits while it penalizes ungrammatical edits, and uses n-grams to capture fluency and grammatical constraints. It does not rely on error categories and is thus a straightforward way to evaluate sequence-to-sequence models (Napoles et al., 2015, 2016).

3. Creating the Test Set

The newly created test set includes common Icelandic spelling and grammar errors, but also errors dependent on context and world knowledge. The first step in creating the test set was text collection, where text sources were searched for particular error categories, and metadata files were created for all collected erroneous documents. The second step was proofreading these documents according to Icelandic spelling and grammar standards, such as the Icelandic Language Council's spelling rules[1] and an official resource on various errors relating to language usage.[2] Only unequivocal errors were corrected and not stylistic ones, so a correction was not made unless the original text was clearly erroneous. Finally, a revision step examined the distribution in error category and data type, and the aforementioned process was repeated to ensure error category and data type distribution. These steps were carried out by a group of three annotators who were all native speakers of Icelandic and had either finished a university degree in Icelandic at the undergraduate level or had significant work experience as professional proof-readers.

The texts to be corrected are sourced from real-world data, i.e. texts which have been written by a third party. Errors are naturally occurring to the greatest extent possible and error examples are of two kinds: *natural examples*, i.e. errors which are found in the original text, and *constructed examples*, i.e. errors that haven't been found in real-world data so a text with the appropriate context is found and it is perturbed so that it becomes erroneous (these instances are much rarer and are recorded in a metadata file for each reviewed text). As mentioned, the test set evaluates the general performance of a spell and grammar checker, while also exercising its context awareness. Therefore, the test set does not consist of single sentences but of whole texts, which are called error documents. Each error document, which can range from being a few sentences to a chapter in an essay, is proofread as a whole.

Two resources were used to search for errors in; a subcorpus of the Icelandic Gigaword Corpus, containing text from news media, both online and written, (Barkarson et al., 2022; Barkarson and Steingrímsson, 2022), along with the Icelandic Common Crawl Corpus (Snæbjarnarson et al., 2022; Miðeind, 2022), which consists of web texts. These corpora reflect modern Icelandic language and a common Icelandic writing style. Variation in written Icelandic is minimal and these resources reflect both relatively formal and informal language use.

The resulting test set is in three parts and contains roughly 380,000 words in total, with more than 9,000 annotations. Texts of type 1 consist of a little less than 200,000 words with around 3,300 annotations, while texts of type 2 consist of just under 150,000 words with roughly 5,000 annotations, and texts of type 3 consist of approximately 30,000 words with around 900 annotations.

[1] https://ritreglur.arnastofnun.is
[2] https://malfar.arnastofnun.is

3.1. Three Types of Test Data

Unlike most test sets for spell and grammar checking, the one discussed here is not annotated in the same way throughout. The test set is in three parts, which are annotated in different ways to facilitate different kinds of evaluation.

Type 1: Labeling only. Error spans in the texts have been marked. The errors are not corrected and individual errors are not labeled further.[3]

Type 2: Correction only. Texts are corrected as a whole, without explicitly marking the span of each error or labeling each error further.[4]

Type 3: Labeling, correction, explanation and severity score. Errors in texts have been marked, corrected and each correction is supported with natural language explanations.[5] Explanations can consist of a few words to a few sentences, e.g. with reference to Icelandic grammar and spelling standards. Providing an explanation to a correction is helpful to users as it gives them nuanced information on the error they made. Additionally, each error is annotated with a severity score on the scale of 1 to 5, 5 being the most severe. Severity scores give information on how important the correction is and the aim of them is to express the potential for reputational impact.

Annotating the documents in different ways allows for different evaluation methods and evaluating different aspects of spell and grammar checkers. Type 1 is the most time-efficient method of creating a test set, as errors are simply marked. This method optimizes the annotator's error labeling throughput, and can thus deliver examples of more text types, vocabularies and error types than the more labor-intensive types. The data resulting from this method can be used to compute error detection accuracy, but it can't be used to evaluate the accuracy of suggested corrections.

Annotating type 2 is less time-efficient than type 1, but it results in more information, i.e. which errors are in the text and how they can be corrected. Although error spans are not explicitly annotated, they can be obtained automatically afterwards by analyzing changes in the document. This method of computing spans can be limiting but it was in part chosen for its simplicity when correcting text, making it possible for annotators to produce more amounts of corrected texts. This data gives us information on error detection accuracy and error correction accuracy, as long as only one correction is available, and can be used to calculate GLEU scores.

Finally, type 3 is a novel kind of test data, providing the most amount of information. Not only does it enable the computation of error detection and error correction accuracy, but it also supplies the reasoning behind the correction and a severity score to the original error. Data can then be stratified by severity and models can be trained on filtered data. This type of data is elemental for evaluating explainable LLMs, in particular LLMs that in addition to correcting, are able to instruct the user on better language use, something that benefits language learners and native speakers alike. Explanations to corrections can be used to train LMs by annotating the training data in an appropriate way so that the model learns to formulate useful explanations to the corrections. These additions to corrections provide useful information when training and evaluating future LLMs.

3.2. Data Format

Texts in the test set are obtained from different sources, which means that they can have different licenses. Where possible, texts published under permissive licenses were used and the resulting test set is published under permissive licenses.

For every original document, at least two files are published, the corrected text or output of the software used to annotate errors, and a metadata file. The metadata file includes information such as text genre, text source and focus error category. Texts from the Icelandic Common Crawl Corpus are published under permissive licenses, so original texts can be published with the test set, which is done as .txt files for all data types. Texts from the Icelandic Gigaword Corpus are, however, published under more restricted licenses, so original texts cannot be published. Instead, for data of type 2, changes to the texts (diffs) are published with a reference to the original text, along with a program which outputs the original text and the corrected one. For data of types 1 and 3, the original accessible document is listed. This approach makes the test data accessible while also making more texts employable when creating the test set.

Corrected data of type 1 is published as JSON Lines files, where each line represents a document. Information shown for each document includes the original text, error spans and their start and end offset. Corrected data of type 2 is published as a .txt file. Error spans are not annotated when the data is created, but they are computed afterwards, showing minimum changes. Finally, corrected data of type 3 is published as .ann files, and information on each document includes an error span's start and end offset, the text included in the span, the corrected version of that text, the severity score and natural language explanation. For more information on the format of

[3] The Doccano annotation tool (Nakayama et al., 2018) is used for this data type.

[4] Any text processing tool can be used when annotating this data.

[5] The Brat annotation tool (Stenetorp et al., 2012) is used for this purpose.

all data types, see the dataset's README file.

3.3. Classifying Documents

Each erroneous document in the dataset is categorized into one or more of five focus error categories, instead of each annotated error within a document being classified. The focus categories were chosen heuristically, based on what kinds of errors we prioritized at this time for evaluating a spell and grammar checker on. Available Icelandic error corpora are descriptive in that they only include errors which are naturally occurring and texts are not chosen for proofreading based on whether they include a certain error. Evaluating spell and grammar checkers on these corpora gives results on the checkers' general performance on Icelandic text, but with the dataset presented in the paper, the aim is to expand the scope of errors that we can evaluate spell and grammar checkers on.

The annotators searched for these error types in extensive text corpora, and corrected the ones found, but if they could not be found, the correct version was found and an error injected into the text, which was then corrected. This process ensures that the dataset consists of these focus error categories. As expected, documents classified as containing a particular error category can contain errors from other categories as well. As a result, we are evaluating a model's performance on a particular type of error and at the same time evaluating its general correction abilities.

The five focus error categories are **idiomatic expressions**, which are Icelandic idioms/phrases with a figurative meaning. People commonly make errors in these idioms; a published language resource is used as a reference for these errors (Halldórsson et al., 2022). **Frequent errors made by Icelandic informants** is used as an umbrella term to comprise various errors which can be found in the texts, e.g. spacing errors, errors relating to punctuation and capitalization, and incorrect cases of nouns, adjectives and pronouns. **Errors relating to context** include inconsistent use of words throughout a text and errors in personal pronouns when they relate to a particular item or person. **Errors relating to cohesion or coherence** are e.g. errors in certain discourse markers, as an example writing 'on the one hand' and then not providing a counterexample, or not using correct pronouns when referring back to previously mentioned objects. Lastly, **semantic analysis** comprises errors which depend on the text's meaning, i.e. real-word errors, errors which cannot be identified and corrected unless the spell and grammar checker has some world knowledge. An example of such an error is 'My ant bought a car'. This sentence is correct with regards to spelling and grammar, but having world knowledge, a proofreader would see that an ant is unlikely to buy a car, so a correction ('aunt') should be provided.

Boundaries between different error categories are not always clear, and ambiguous errors arose when the test set was created. An example of this is the aforementioned error 'My ant bought a car', where the 'ant' error can be considered an error due to semantic analysis or as a typographical error. Both classifications can be reasoned, and edge cases were discussed in detail amongst the annotators before reaching a conclusion on how to classify them.

3.4. Inter-Annotator Agreement

To measure inter-annotator agreement on the data, we prepared 168 examples for evaluation where an annotator had to indicate preference for an original sentence or a corrected sentence. The ordering of examples was random, i.e., the annotator was blinded towards which example was the original and which one was corrected. Four participants, separate from the test set's annotators, performed the evaluation on all examples. They all had either finished a university degree in Icelandic at the undergraduate level or had significant work experience as professional proofreaders. On average, the corrected sentences were preferred in 92.3% of cases (ranging from 87.5% to 94.6% for the annotators). We computed inter-annotator agreement using Krippendorff's Alpha (Krippendorff, 2018) and the result was a score of 0.829, indicating almost perfect reliability.

4. Discussion

Creating this test data as described above, using the resources mentioned, has the possible limitation of underlying texts being used for training LLMs, since some of them are sourced from the internet. This is hard to avoid, as we need a large corpus in order to find naturally occurring errors. On the other hand, in most cases, it is the erroneous version that is in the training data, not the one corrected by our experts.

As part of future work, an LLM will be fine-tuned on the spell and grammar checking task for Icelandic. Following this is possible work on enhancing text beyond correcting explicit errors, e.g. improving text fluency and making stylistic changes to better conform to a particular register. Changes to be made can be less distinct when it comes to these categories, so which guidelines should be followed would have to be considered.

5. Conclusion

We have presented a new test set for evaluating automatic spell and grammar checkers of different kinds, in particular large language mod-

els. The test set is manually annotated for Icelandic spelling and grammar errors with a focus on context-dependent errors. The data is annotated in three different ways: with span-marking, with corrections and with natural language explanations of corrections and severity scores. Explanations of corrections and error severity scores are a novel addition to test data, particularly intended for evaluating LLMs. The test set can be used to evaluate current and future spell and grammar checking systems and is published under a permissive license (Símonarson et al., 2023).

6. Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019–2023. The programme, which is managed and coordinated by Almannarómur (https://almannaromur.is/), is funded by the Icelandic Ministry of Education, Science and Culture.
We would like to thank the anonymous reviewers for their valuable feedback.

7. Bibliographical References

Þórunn Arnardóttir, Isidora Glisic, Annika Simonsen, Lilja Stefánsdóttir, and Anton Ingason. 2022. Error corpora for different informant groups: Annotating and analyzing texts from L2 speakers, people with dyslexia and children. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 245–252, New Delhi, India. Association for Computational Linguistics.

Þórunn Arnardóttir, Xindan Xu, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir, and Anton Karl Ingason. 2021. Creating an error corpus: Annotation and applicability. In *CLARIN Annual Conference Proceedings*, pages 59–63.

Starkaður Barkarson, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022. Evolving large text corpora: Four versions of the Icelandic Gigaword corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2371–2381, Marseille, France. European Language Resources Association.

Marie Bexte, Ronja Laarmann-Quante, Andrea Horbach, and Torsten Zesch. 2022. LeSpell - a multi-lingual benchmark corpus of spelling errors to develop spellchecking methods for learner language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 697–706, Marseille, France. European Language Resources Association.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is ChatGPT a highly fluent grammatical error correction system? A comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Svanhvít Lilja Ingólfsdóttir, Pétur Ragnarsson, Haukur Jónsson, Haukur Símonarson, Vilhjálmur Þorsteinsson, and Vésteinn Snæbjarnarson. 2023. Byte-level grammatical error correction using synthetic and curated corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 7299–7316, Toronto, Canada. Association for Computational Linguistics.

Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao. 2017. A nested attention neural hybrid model for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 753–762, Vancouver, Canada. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana. Association for Computational Linguistics.

Anisia Katinskaia, Maria Lebedeva, Jue Hou, and Roman Yangarber. 2022. Semi-automatically annotated learner corpus for Russian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 832–839, Marseille, France. European Language Resources Association.

Katerina Korre and John Pavlopoulos. 2022. Enriching grammatical error correction resources for Modern Greek. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4984–4991, Marseille, France. European Language Resources Association.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023. On the (in)effectiveness of large language models for Chinese text correction. *arXiv preprint arXiv:2307.09007*.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada. Association for Computational Linguistics.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from `https://github.com/doccano/doccano`.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2016. GLEU without tuning. *arXiv preprint arXiv:1605.02592*.

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.

Anna Nikulásdóttir, Þórunn Arnardóttir, Starkaður Barkarson, Jón Guðnason, Þorsteinn Gunnarsson, Anton Ingason, Haukur Jónsson, Hrafn Loftsson, Hulda Óladóttir, Eiríkur Rögnvaldsson, Einar Sigurðsson, Atli Sigurgeirsson, Vésteinn Snæbjarnarson, Steinþór Steingrímsson, and Gunnar Örnólfsson. 2022. Help yourself from the buffet: National language technology infrastructure initiative on CLARIN-IS. In *Selected Papers from the CLARIN Annual Conference 2021*. Linköping Electronic Conference Proceedings.

Hulda Óladóttir, Þórunn Arnardóttir, Anton Ingason, and Vilhjálmur Þorsteinsson. 2022. Developing a spell and grammar checker for Icelandic using an error corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4644–4653, Marseille, France. European Language Resources Association.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman,

Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *arXiv preprint https://arxiv.org/abs/2303.08774*.

Maria Carolina Penteado and Fábio Perez. 2023. Evaluating GPT-3.5 and GPT-4 on grammatical error correction for Brazilian Portuguese. *arXiv preprint arXiv:2306.15788*.

Fanyi Qu and Yunfang Wu. 2023. Evaluating the capability of large-scale language models on Chinese grammatical error correction task. *arXiv preprint arXiv:2307.03972*.

Georg Rehm and Andy Way, editors. 2023. *European Language Equality - A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjálmur Þorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the*

Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 102–107, Avignon, France. Association for Computational Linguistics.

Yujin Takahashi, Masahiro Kaneko, Masato Mita, and Mamoru Komachi. 2022. ProQE: Proficiency-wise quality estimation dataset for grammatical error correction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5994–6000, Marseille, France. European Language Resources Association.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Baoxin Wang, Xingyi Duan, Dayong Wu, Wanxiang Che, Zhigang Chen, and Guoping Hu. 2022. CCTC: A cross-sentence Chinese text correction dataset for native speakers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3331–3341, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yu Wang, Yuelin Wang, Jie Liu, and Zhuo Liu. 2020. A comprehensive survey of grammar error correction. *arXiv preprint arXiv:2005.06600*.

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*.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California. Association for Computational Linguistics.

8. Language Resource References

Starkaður Barkarson and Steinþór Steingrímsson. 2022. IGC-news2-22.10 (annotated version). CLARIN-IS.

Björn Halldórsson, Árni Davíð Magnússon, Finnur Ágúst Ingimundarson, Einar Freyr Sigurðsson, Steinþór Steingrímsson, Halldóra Jónsdóttir, and Þórdís Úlfarsdóttir. 2022. Idiomatic expressions (Icelandic and English) 22.09. CLARIN-IS.

Anton Karl Ingason, Þórunn Arnardóttir, Lilja Björk Stefánsdóttir, and Xindan Xu. 2021a. The Icelandic child language error corpus (IceCLEC) version 1.1. CLARIN-IS.

Anton Karl Ingason, Þórunn Arnardóttir, Lilja Björk Stefánsdóttir, Xindan Xu, Dagbjört Guðmundsdóttir, and Isidora Glišić. 2022a. The Icelandic dyslexia error corpus 1.2 (22.10). CLARIN-IS.

Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, and Xindan Xu. 2021b. Icelandic error corpus (IceEC) version 1.1. CLARIN-IS.

Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, Xindan Xu, Isidora Glišić, and Dagbjört Guðmundsdóttir. 2022b. The Icelandic L2 error corpus (IceL2EC) 1.3 (22.10). CLARIN-IS.

Miðeind. 2022. Icelandic common crawl corpus (IC3). Hugging Face.

Haukur Barri Símonarson, Svanhvít Lilja Ingólfsdóttir, Þórunn Arnardóttir, Dagbjört Guðmundsdóttir, Ella María Georgsdóttir, and Guðrún Lilja Friðjónsdóttir. 2023. Grammatical error correction test set. CLARIN-IS.