Playing by the Rules: A Benchmark Set for Standardized Icelandic Orthography

Bjarki Ármannsson, Hinrik Hafsteinsson, Jóhannes B. Sigtryggsson, Atli Jasonarson, Einar Freyr Sigurðsson, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
bjarki.armannsson@arnastofnun.is, hinhaf@hi.is,
{johannes.b.sigtryggsson, atli.jasonarson,
einar.freyr.sigurdsson, steinthor.steingrimsson}@arnastofnun.is

Abstract

We present the Icelandic Standardization Benchmark Set: Spelling and Punctuation (IceStaBS:SP), a dataset designed to provide standardized text examples for Icelandic orthography. The dataset includes non-standard orthography examples and their standardized counterparts, along with detailed explanations based on the official Icelandic spelling rules. IceStaBS:SP aims to support the development and evaluation of automatic spell and grammar checkers, particularly in an educational setting. We evaluate various spell and grammar checkers using IceStaBS:SP, demonstrating its utility as a benchmarking tool and highlighting areas for future improvement.

1 Introduction

Digital language infrastructure, not least for spell and grammar checking, is a productive and growing field within Icelandic Language Technology. Although various datasets have been produced, which in turn have been used to develop and improve spell and grammar checking software, there is a lack of datasets which provide a 1:1 mapping between spelling errors and formalized rules regarding standard orthography (spelling rules).

In this paper, we present the Icelandic Standardization Benchmark Set: Spelling and Punctuation (IceStaBS:SP, Ármannsson et al. 2024), a dataset of examples of text standardization along with thorough explanations of how and why text has been altered. The dataset is based on the official spelling rules for Icelandic.[1] Our goal is to provide a standardized benchmark for evaluating the performance of spell and grammar checkers, thereby contributing to the improvement of digital language tools for Icelandic.

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

The paper is structured as follows: Section 2 provides an overview of related work in the field of Icelandic spell and grammar checking, most importantly existing datasets. Section 3 describes the structure of the IceStaBS:SP dataset and the methodology behind it. Section 4 outlines the evaluation experiment we performed to gauge the efficacy of the dataset as a benchmarking tool for orthography. Section 5 presents the results of the evaluation, Section 6 discusses the limitations of our approach and Section 7 concludes the paper with a discussion of the implications of our findings and suggestions for future work.

2 Related Work

The most comprehensive single dataset in the field of spell and grammar checking for Icelandic is the Icelandic Error Corpus (IEC, Arnardóttir et al. 2021) and its subsidiary corpus for errors made by L2 speakers (Glišić and Ingason, 2021). It uses a fine-grained error categorization system and has been used for training and evaluating spell and grammar checkers, specifically the rule-based GreynirCorrect (Óladóttir et al., 2022).

The Grammatical Error Correction Test Set (GECTS, Arnardóttir et al. 2024b), a hand-annotated dataset of Icelandic text with various spelling and grammatical errors, is annotated on the document level as opposed to the IEC, where each individual error is annotated. This, along with a more general error categorization system, makes it more suitable for evaluating recent sequence-to-sequence error correction models by testing the models' context awareness on larger texts.

3 Suggesting Standardized Orthography

We present the Icelandic Standardization Benchmark Set: Spelling and Punctuation (IceStaBS:-SP), a dataset of text examples containing non-

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 28–36

standard orthography and their standardized counterparts. Each item in our set corresponds to an entry in the official spelling rules for Icelandic, published by the Icelandic Language Council and applied in Icelandic schools. Each item contains three examples of standardized text, along with thorough explanations of how and why each text has been altered. The dataset is meant to serve as a key component in the development of automatic spell checking in an educational setting, providing handcrafted explanations which can be expanded or used for instruction tuning.

Both the text examples showing non-standard orthography and the additional explanations are constructed and reviewed by the authors of this paper, all of whom have a background in Icelandic linguistics and one of whom is one of the authors of the most recent version of the official spelling rules. The text examples each show exactly one non-standard text feature in order to clearly demonstrate the applicable standardization and in order to check whether that feature has been correctly captured by a spell-checking system. Depending on the orthographic issue being demonstrated, the text examples range from very short and simple sentences to short paragraphs, e.g. to display the prescribed use of punctuation between whole sentences. They are mostly synthetic (and partly based on the examples included in the publication of the spelling rules themselves) but where possible, we have extracted real-world examples from the IEC using the error codes in that corpus. This authentic approach was, however, limited by the need to include only one example of non-standard orthography in each example, so some of those examples have been slightly altered.

The official spelling rules consist of 33 main chapters and numerous subchapters. Some subchapters, as well as all of chapters 30 and 33, are ignored in our set as they are not applicable in the context of automatic correction of spelling and grammar (e.g. some contain only general discussion of phenomena, rather than concrete examples). In other cases, subchapters had to be split into further subsections for our purposes as they dealt with multiple distinct features. These are marked with "(a)", "(b)", etc. in IceStaBS:SP. In this way, we define a total of 247 rules over 31 chapters.

For each of these 247 rules, our dataset contains an entry labeled with a distinct number and consisting of the following parts:

- **Short suggestion:** A suggested format for displaying a correction made by a spellchecker, containing a brief summary of the applicable spelling rule.
- **Long suggestion:** A more detailed description of the applicable rule, complete with a URL to the relevant chapter of the official spelling rules (in a few cases, links to multiple rules are included).
- **Examples:** Three examples of a short text containing the relevant issue, which show a potential correction in a hypothetical spell-correction interface according to the 'short suggestion' format. The proposed changes are shown both in isolation and in the context of the whole text.[2]
- **Error Code:** The relevant error code in the IEC.
- **URL:** The URL of the relevant section of the official spelling rules.

To illustrate how this information is structured in the IceStaBS:SP dataset, the entry for rule 1.2.1 (a) is shown in Figure 1.

The aim of the suggestions and explanations in our set is to provide further assistance to potential future users of an automatic spellchecker, not least young people and second language learners of Icelandic. Therefore, we try to keep our explanations accessible to the average speaker, with as little linguistic terminology as possible (especially in the short suggestions).

To as great an extent as possible, we also try to include helpful generalizations in the short suggestion format as opposed to only word-specific corrections. An example would be *<villa> á líklega að vera með stórum staf, <leiðrétt>, þar sem það er örnefni* '<error> should likely be written with a capital initial letter, <correction>, as it is a place name', rather than simply '<error> should likely be written with a capital initial letter'. This is sometimes made difficult, however, by rules that can apply to many different scenarios or are simply too complex to sum up in one short sentence.

[2]In a few cases, these entries will be identical. This is mostly in the case of punctuation, e.g. where rules on appropriate marking of a subclause need to take into account the whole sentence.

```
"1.2.1 (a)": {
    "short_suggestion": "<villa> á líklega að vera með stórum staf, <leiðrétt>, þar sem það kemur
                         á eftir punkti.",
    "long_suggestion": "Stór stafur er alltaf ritaður í upphafi máls og í nýrri málsgrein
                        á eftir punkti.
                         Sjá ritreglu 1.2.1 (https://ritreglur.arnastofnun.is/#1.2.1).",
    "examples": {
        "1": {
            "original_sentence": "Afi og amma ætla að koma í heimsókn. þau koma bráðum.",
            "standardized_sentence": "Afi og amma ætla að koma í heimsókn. Þau koma bráðum.",
            "suggestion": "<þau> á líklega að vera með stórum staf, <Þau>, þar sem það kemur
                           á eftir punkti.",
            "original_part": "þau",
            "standardized_part": "Þau"
        },
        "2": {
            "original_sentence": "Ráðgert er að nýtt hús rísi í vor. vinnan við það er þó ekki hafin.",
            "standardized_sentence": "Ráðgert er að nýtt hús rísi í vor. Vinnan við það er þó ekki hafin.",
            "suggestion": "<vinnan> á líklega að vera með stórum staf, <Vinnan>, þar sem það kemur
                           á eftir punkti.",
            "original_part": "vinnan",
            "standardized_part": "Vinnan"
        },
        "3": {
            "original_sentence": "Margt skiptir máli þegar skáldsögur eru skrifaðar. málfar er t.d.
                                  mikilvægur þáttur.",
            "standardized_sentence": "Margt skiptir máli þegar skáldsögur eru skrifaðar. Málfar er t.d.
                                      mikilvægur þáttur.",
            "suggestion": "<málfar> á líklega að vera með stórum staf, <Málfar>, þar sem það kemur
                           á eftir punkti.",
            "original_part": "málfar",
            "standardized_part": "Málfar"
        }
    },
    "error_code": "lower4upper-initial",
    "ritreglur_url": "https://ritreglur.arnastofnun.is/#/1.2.1 (a)"
}
```


Figure 1: JSON structure of the IceStaBS:SP dataset, showing the entry for rule 1.2.1 (a), which deals with capitalization after a full stop. The text in the 'short suggestion' slot says: '<error> should probably be capitalized, <correction>, as it follows a full stop.' The text in the 'long suggestion' slot says: 'A capital letter is always used at the start of a text and the beginning of a new sentence following a full stop. See spelling rule 1.2.1 [. . .].' Examples 1–3 then show text in Icelandic where the start of a sentence has not been capitalized, with suggested corrections in the 'suggestion' slot presented according to the 'short suggestion' format.

4 Applying IceStaBS:SP in Evaluation

To gauge the efficacy of the IceStaBS:SP dataset as a benchmarking tool for orthography, we performed an evaluation experiment, where various spell and grammar checkers for Icelandic were applied on our data and then evaluated statistically. This serves two purposes.

Firstly, it allows us to evaluate the performance of these tools on a standardized dataset, which can be used to compare the tools to each other and, preferably, to other benchmark sets. Secondly, we standardize our methods for evaluating correction tools on our benchmark set. The source code of our evaluation methods is then made available on GitHub[3] for others to use on new tools, as well as the output of the tools we use in our evaluation.

[3] https://github.com/stofnun-arna-magnussonar/IceStabs-eval

4.1 Tools Evaluated

We intend our dataset to be applicable to any tool which corrects errors in Icelandic text. With this in mind, we selected 10 tools and models to test. These include commercial and open-source software, with a broad range of effectiveness, from state-of-the-art to baseline tools.

Our first focus are tools which can be run programmatically. These were:

Byte-Level Neural Error Correction Model for Icelandic (Ingólfsdóttir et al., 2023): A fine-tuned ByT5-base Transformer designed for error correction in Icelandic text. It functions similarly to a machine translation model, converting erroneous Icelandic into correct Icelandic. We evaluate three versions of this tool, each representing a successive update: `22-09`, `23-12`, and `24-03`.

GreynirCorrect (Óladóttir et al., 2022): A rule-

based spell and grammar checker for Icelandic. The tool is based on Greynir (Þorsteinsson et al., 2019), a syntactic parser for Icelandic. We evaluate the most recent version of this tool: version `4.0.0`.[4]

Icelandic GPT-SW3 for Spell and Grammar Checking (Arnardóttir et al., 2024a): A GPT-SW3 (Ekgren et al., 2022, 2024) model, fine-tuned on Icelandic and particularly on the task of spell and grammar checking. The experimental setup we use is identical to the example given in the model's HuggingFace repository.[5]

Skrambi: A closed-source rule-based spell checker for Icelandic.[6]

In addition to the tools which can be run programmatically, we evaluated four 'manual' tools, i.e., tools which are first and foremost accessible through an end-user platform of some kind. These were:

Hunspell: An open-source spell checker and morphological analyzer. The Icelandic language rules[7] for Hunspell are accessible via LibreOffice, where Hunspell is the standard spell checker.

Google Docs Spelling and Grammar check: Built-in spell and grammar checker of Google Docs.[8]

Microsoft Editor: Built-in spell and grammar checker of Microsoft Word.[8]

Ritvilluvörnin Púki: Proprietary spell and grammar checker, specifically for Icelandic text.[9]

With this in mind, the total number of tools and individual correction models we evaluate is 10. Half of these (5) are developed by Miðeind,[10] a private language technology company.

In our evaluation experiment, each tool is given a simplified label, which we will use to refer to them in the following sections. An alphabetic overview of these labels is as follows:

1. `ByT5 (22-09)`
2. `ByT5 (23-12)`
3. `ByT5 (24-03)`
4. `Google Docs`
5. `GreynirCorrect`
6. `Hunspell`
7. `Ice-GPT-SW3`
8. `MS Word`
9. `Púki`
10. `Skrambi`

4.2 Evaluation Metrics

We define three main metrics which can be used to evaluate the performance of a spell and grammar checker on our dataset:

Sentence-level accuracy: Direct comparison between output sentences and standardized versions. A sentence is considered correct if the output is identical to the standardized sentence.

Token-level $F_{0.5}$ score: An F-score metric modified for spell and grammar correction. $F_{0.5}$ is a weighted average of precision and recall, where precision is given twice the weight of the recall. It is included in the ERRANT toolkit (Bryant et al., 2017) and was used in the CoNLL-2014 shared task (Ng et al., 2014).

GLEU score: A modified version of the BLEU score. BLEU is used to evaluate the quality of machine translation, while GLEU is used to evaluate the quality of spell and grammar correction. It is especially well suited for evaluating sequence-to-sequence models, as it does not rely on error categories for evaluation (Napoles et al., 2015, 2016).

As the tools we evaluate are technically and functionally diverse, it may be inferred that a given metric may suit one tool better than another. This is up to analysis, but in our overall evaluation structure, we use all three metrics to evaluate all tools.

5 Results

We evaluate the performance of the tools on the IceStaBS:SP dataset using the three metrics described above. The results are shown respectively in Figures 2, 3, and 4.

5.1 Performance Per Tool

The tool with both the highest proportion of correct sentences, as shown in Figure 2, and the highest $F_{0.5}$ score, as shown in Figure 3, is Miðeind's ByT5 (23-12) with 46.42% sentence accuracy and a token-level $F_{0.5}$ score of 0.70.

[4] `https://github.com/mideind/GreynirCorrect`

[5] `https://huggingface.co/mideind/icelandic-gpt-sw3-6.7b-gec/blob/main/handler.py`

[6] `https://skrambi.arnastofnun.is`

[7] `https://github.com/nifgraup/hunspell-is`

[8] Publically available versions as of October 27, 2024.

[9] `https://puki.is`

[10] `https://mideind.is`

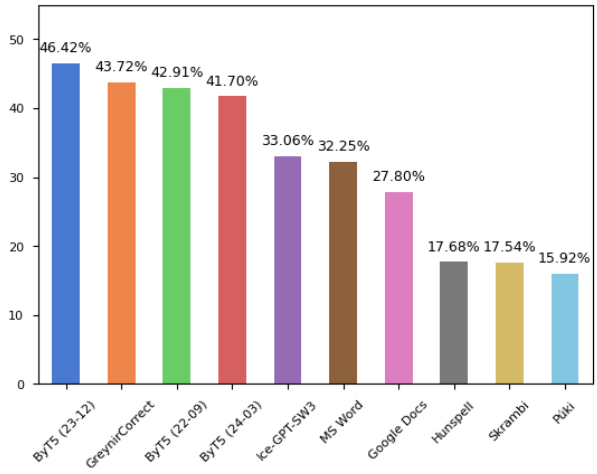

Figure 2: Sentence-level accuracy of the tools evaluated.

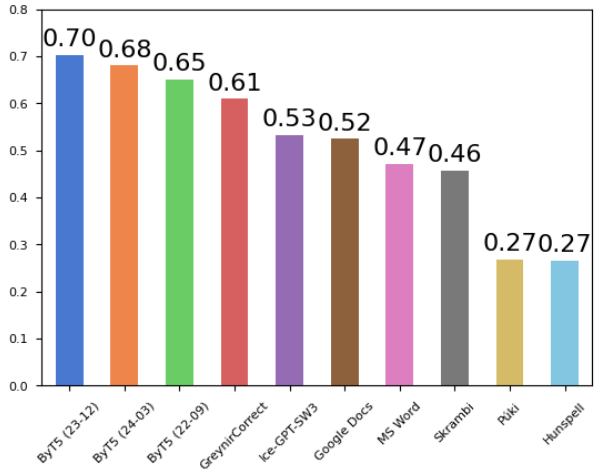

Figure 3: Token-level $F_{0.5}$ scores of the tools evaluated.

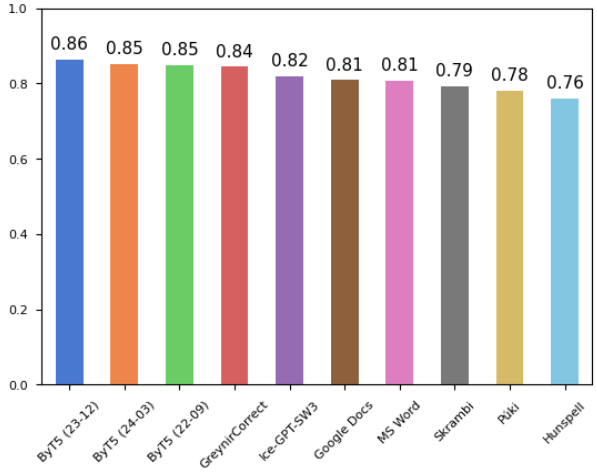

Figure 4: Dataset-level GLEU scores of the tools evaluated.

One possible limitation of the results described here is that not all spell-checking software has equal coverage when it comes to our predefined rule chapters. In the case of MS Word, Google Docs and Púki, various errors are not handled by the spell-checking features of the platform, but the respective autocorrect functionality of the platform. This particular issue is beyond the scope of our current evaluation but will hopefully be controlled for in a future iteration.

We see substantial variance between the highest scoring tools and the lowest. This is especially interesting when real-world integration and use are taken into account. Púki, the widely used spell-checking tool for Icelandic (originally released in 1987 and iterated upon since then), achieves the lowest scores on our sentence correctness and token-level $F_{0.5}$ score metrics.

The leaders of the evaluation metrics are the Miðeind ByT5 models, followed closely by GreynirCorrect. On the one hand, it is interesting that of the three ByT5 models, the newest iteration (24-03) underperforms compared to the previous one (23-12). On the other hand, all the (comparably lightweight) ByT5 models, along with the rule-based GreynirCorrect, outperform the compute-heavy Ice-GPT-SW3 model.

As shown in Figure 4, the three ByT5 models achieve the highest GLEU scores. It should be noted that of the tools we evaluate, the ByT5 (24-03) and Ice-GPT-SW3 models have previously published GLEU scores: ByT5 (24-03) is reported to achieve GLEU scores of 0.90 and 0.91 when evaluated on the GECTS and IEC datasets, respectively (Ingólfsdóttir et al., 2023). The GLEU score for the Ice-GPT-SW3 when evaluated on the GECTS is 0.93 (Arnardóttir et al., 2024a). These numbers are different from our results, which may reflect inherent differences in IceStaBS:SP, compared to the GECTS and IEC datasets. This is not totally unexpected, as the other two datasets are corpora, which IceStaBS:SP is not. Further analysis will shed light on these differences.

5.2 Performance Per Rule Chapter

As is to be expected, as the phenomena dealt with in some chapters are more common or straightforward than in others, there is considerable variance in tool performance across different chapters of the spelling rules. The highest scores recorded for each chapter are shown in Tables 1 and 2, in

terms of $F_{0.5}$ score and sentence accuracy, respectively. In some cases, the best-performing tools correct each example in a chapter exactly as intended. These include chapters 9 and 18, both of which have only three example texts in our dataset and deal with common and fairly straightforward issues (chapter 9 concerns words such as *hvar* ('where') that are spelt with *hv* and not *kv*, despite the *h* invariably being pronounced [k^h] and not [h] by most speakers, and chapter 18 deals with double consonant stems).

More pleasantly surprising is the excellent maximal performance achieved on chapters 6 (11 out of 12 examples correct), 16 (11 out of 12 correct) and especially 4 (21 out of 21 examples correct). The top-scoring model for the last of these chapters proved to be the rule-based and computationally light Skrambi, which overall placed 8th out of the ten models in terms of token-level $F_{0.5}$ score, behind all neural models. It is worth noting that chapter 4 concerns the spelling of vowels before the consonants clusters *ng* and *nk*, where letters that typically are used to represent monophthongs are pronounced as diphthongs (e.g. *banki* ('bank') instead of **bánki*, even though the relevant sound, [au], is almost always represented with *á* and not *a*) but the opposite can also occur without any cues in pronunciation in some exceptions (e.g. *jánka* ('agree'), derived from *já* ('yes'), or *rángirni* ('greed'), a compound formed by *rán* ('robbery') and *girni* ('desire')). This is an example of a scenario that seems to lend itself better to models that are rule-based or include hard-coded exceptions, as opposed to neural models which might possibly be thrown off the trail of the overarching rule by exceptions found in the training data.

On the other hand, for chapters 23 (which covers semicolons) and 27 (which covers parentheses and square brackets), not a single tool corrected a single example in accordance with the spelling rules. Both those rules fall under the punctuation part of the spelling rules, which somewhat predictably yields generally worse results than the spelling portion (chapters 1 through 20). After all, rules on punctuation often depend on some fairly abstract semantic features (e.g. from rule 23.1: 'A semicolon represents a stronger break in a text than a comma but a lesser break than a full stop') and deviations from the standard do not result in non-words, as deviations from spelling rules might.

Ch.	Best Tool	$F_{0.5}$	No. Ex.
1	greynir	0.46	153
2	byt5-23-12	0.60	60
3	skrambi	0.83	12
4	skrambi	1	21
5	word	0.56	75
6	google	0.96	12
7	greynir	0.66	21
8	greynir	0.74	39
9	byt5-22-09	1	3
10	greynir	0.62	21
11	byt5-22-09	0.66	3
12	byt5-23-12	0.85	30
13	google	0.50	12
14	byt5-24-03	0.8	30
15	byt5-22-09	0.57	36
16	greynir	0.91	12
17	hunspell	0.67	9
18	byt5-22-09	1	3
19	google	0.74	21
20	hunspell	1	6
21	ice-gpt-sw3	0.28	42
22	byt5-23-12	0.54	24
23	**None**	0	3
24	byt5-24-03	0.16	6
25	byt5-22-09	0.66	3
26	ice-gpt-sw3	0.38	33
27	**None**	0	6
28	byt5-22-09	0.5	6
29	byt5-22-09	0.27	18
30	ice-gpt-sw3	0.17	9
31	byt5-23-12	0.33	12

Table 1: $F_{0.5}$ score Leaderboard for IceStabs:SP Evaluation, for each chapter in the spelling rules.

6 Limitations

There are various aspects of the IceStaBS:SP dataset that could be improved in future iterations. These range from superficial to inherent issues, the solutions to which will need further work and discussion.

As shown in Figure 5, even though the IceStabs:SP data is organized into 31 distinct chapters (reflecting the 33 chapters of the source material), the distribution of examples across these chapters is not uniform. This is due to the fact that some chapters cover more common and straightforward spelling rules, while others deal with more com-

Ch.	Best Tool	Score	Total	Ratio
1	byt5-23-12	68	153	44.44%
2	byt5-22-09	34	60	56.66%
3	skrambi	10	12	83.33%
4	skrambi	21	21	100%
5	word	40	75	53.33%
6	byt5-23-12	11	12	91.66%
7	greynir	14	21	66.66%
8	greynir	28	39	71.79%
9	byt5-22-09	3	3	100%
10	google	12	21	57.14%
11	byt5-22-09	2	3	66.66%
12	byt5-22-09	25	30	83.33%
13	greynir	6	12	50%
14	byt5-23-12	23	30	76.66%
15	byt5-22-09	20	36	55.55%
16	greynir	11	12	91.66%
17	hunspell	5	9	55.55%
18	byt5-22-09	3	3	100%
19	google	15	21	71.42%
20	hunspell	6	6	100%
21	ice-gpt-sw3	12	42	28.57%
22	byt5-23-12	13	24	54.16%
23	**None**	0	3	**N/A**
24	byt5-24-03	1	6	16.66%
25	byt5-22-09	2	3	66.66%
26	ice-gpt-sw3	12	33	36.36%
27	**None**	0	6	**N/A**
28	byt5-22-09	3	6	50
29	byt5-22-09	5	18	27.77%
30	greynir	2	9	22.22%
31	byt5-23-12	4	12	33.33%

Table 2: Sentence-level Accuracy Leaderboard for IceStabs:SP Evaluation, for each chapter in the spelling rules.

plex, subjective or less frequent issues. As a result, the dataset contains a larger number of examples for the more common rules, which may skew the evaluation results towards these chapters. In short, not all chapters of the Icelandic spelling rules are created equal.

Chapters 1, 2, and 5 have significantly more entries than the other chapters in the dataset, with chapter 1 (use of upper and lower case letters at the beginnings of words) being particularly prominent. This discrepancy is due to the fact that these chapters cover fundamental and frequently encountered spelling rules in Icelandic orthography.

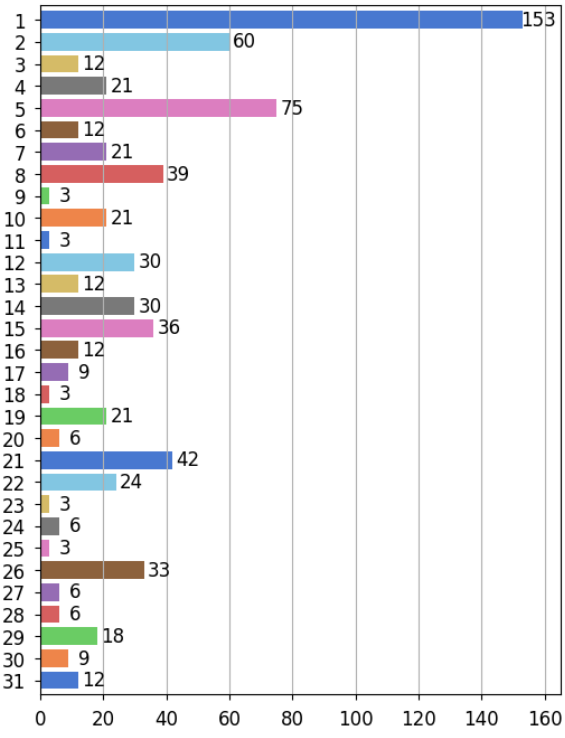

Figure 5: Number of example sentences per main chapter in the IceStaBS:SP dataset.

On the opposite end of this spectrum are chapters 9, 11, 18, 23 and 25, which all have a single rule entry each (giving 3 examples per chapter in the overview in Figure 5). Even though the dataset structure and evaluation procedure treats these chapters as equal to the others, they are not equal in terms of the number of examples.

Currently, the IceStaBS:SP dataset only allows for a single standardized suggestion for each example. This means that the dataset does not account for the possibility of multiple correct solutions to a given spelling or grammar issue. As there are sometimes more than one correct way to write something according to the spelling rules, some entries in our dataset *should* allow for multiple possible correct alterations. An example would be some non-standard way of writing a specific time, which could be corrected to e.g. '2.30' or '2:30' as both a full stop and a colon are possible ways of separating hours from minutes, according to spelling rules 22.5 and 29.5, respectively. Even though the number of these occurrences is low (variation is found in about 20 rules out of 247), this is a limitation that will be addressed in future iterations of the dataset.

7 Conclusions and Future Work

We have presented the IceStaBS:SP dataset, a comprehensive benchmark set for Icelandic spelling and punctuation. The dataset is based on the official spelling rules for Icelandic and provides standardized suggestions for a wide range of spelling and punctuation issues. As such, it is the first of its kind for Icelandic.

We have evaluated the performance of ten spell and grammar checkers on the dataset, using three main metrics: sentence-level accuracy, token-level $F_{0.5}$ score, and GLEU score. The results are broadly in line with expected performance, which is encouraging for the utility of the dataset as a benchmarking tool.

Further work is needed to address limitations in the dataset and explore additional evaluation metrics to provide a more comprehensive assessment of spell and grammar checkers for Icelandic.

Acknowledgments

We would like to thank the NoDaLiDa/Baltic-HLT 2025 organizers for the assistance and communication while working on this submission and three anonymous reviewers for helpful feedback. In addition to the authors of this paper, Finnur Ágúst Ingimundarson is one of the authors of the dataset described here. This work was financed by the Icelandic Ministry of Culture and Business Affairs as part of the Language Technology Programme for Icelandic.

References

Bjarki Ármannsson, Hinrik Hafsteinsson, Jóhannes B. Sigtryggsson, Atli Jasonarson, Finnur Ágúst Ingimundarson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2024. Icelandic Standardization Benchmark Set: Spelling and Punctuation 24.10. CLARIN-IS.

Þórunn Arnardóttir, Svanhvít Lilja Ingólfsdóttir, Garðar Ingvarsson Juto, Haukur Barri Símonarson, Hafsteinn Einarsson, Anton Karl Ingason, and Vilhjálmur Þorsteinsson. 2024a. Icelandic GPT-SW3 for spell and grammar checking. CLARIN-IS.

Þórunn Arnardóttir, Svanhvít Lilja Ingólfsdóttir, Haukur Barri Símonarson, Hafsteinn Einarsson, Anton Karl Ingason, and Vilhjálmur Þorsteinsson. 2024b. Beyond error categories: A contextual approach of evaluating emerging spell and grammar checkers. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 45–52, Torino, Italia. ELRA and ICCL.

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

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.

Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson, and Magnus Sahlgren. 2022. Lessons learned from GPT-SW3: Building the first large-scale generative language model for Swedish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3509–3518, Marseille, France. European Language Resources Association.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Alice Heiman, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Nordic languages.

Isidora Glišić and Anton Karl Ingason. 2021. The nature of Icelandic as a second language: An insight from the learner error corpus for Icelandic. In *CLARIN Annual Conference*, pages 26–30. Virtual edition.

Svanhvít Lilja Ingólfsdóttir, Pétur Orri Ragnarsson, Haukur Páll Jónsson, Haukur Barri 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.

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.

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.

Hulda Óladóttir, Þórunn Arnardóttir, Anton Karl 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.

Vilhjálmur Þorsteinsson, Hulda Óladóttir, and Hrafn Loftsson. 2019. A wide-coverage context-free grammar for Icelandic and an accompanying parsing system. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1397–1404, Varna, Bulgaria. INCOMA Ltd.