A Warm Start and a Clean Crawled Corpus -
A Recipe for Good Language Models

Vésteinn Snæbjarnarson[1,2,*], **Haukur Barri Símonarson**[1,2,*],
Pétur Orri Ragnarsson[1], **Svanhvít Lilja Ingólfsdóttir**[1], **Haukur Páll Jónsson**[1],
Vilhjálmur Þorsteinsson[1], **Hafsteinn Einarsson**[2]
[1] Miðeind ehf., [2] University of Iceland
{vesteinn, haukur, petur, svanhvit, haukurpj, vt}@mideind.is, hafsteinne@hi.is

Abstract

We train several language models for Icelandic, including IceBERT, that achieve state-of-the-art performance in a variety of downstream tasks, including part-of-speech tagging, named entity recognition, grammatical error detection and constituency parsing. To train the models we introduce a new corpus of Icelandic text, the Icelandic Common Crawl Corpus (IC3), a collection of high quality texts found online by targeting the Icelandic top-level-domain `.is`. Several other public data sources are also collected for a total of 16GB of Icelandic text. To enhance the evaluation of model performance and to raise the bar in baselines for Icelandic, we manually translate and adapt the WinoGrande commonsense reasoning dataset. Through these efforts we demonstrate that a properly cleaned crawled corpus is sufficient to achieve state-of-the-art results in NLP applications for low to medium resource languages, by comparison with models trained on a curated corpus. We further show that initializing models using existing multilingual models can lead to state-of-the-art results for some downstream tasks.

Keywords: language model, Icelandic, IceBERT, corpus, part of speech, named entity recognition, parsing, co-reference resolution, natural language understanding

1. Introduction

The Government of Iceland recently launched an initiative to improve the state of Icelandic language resources and language technology (Nikulásdóttir et al., 2020). This comprehensive program has its roots in the historical focus on protecting the Icelandic language (Kristinsson, 2018), and, as a result, work has been ongoing to build and enhance said resources. This effort is gradually pushing Icelandic from being a low-resource language to a medium-resource one[1]. Still, and apart from large monolingual corpora (Steingrímsson et al., 2018), many important types of resources are lacking in comparison with major languages such as English.

Parallel to the development of the Icelandic language technology program, language technology world-wide has been progressing at a fast and accelerating pace (Bommasani et al., 2021). Pre-trained neural language models based on Transformers (Vaswani et al., 2017) have shown impressive results when adapted for a variety of classification and text generation tasks. Such models are now applied widely across industries and modalities.

Large monolingual language models such as BERT (Devlin et al., 2019), BART (Lewis et al., 2020) and GPT-2 (Radford et al., 2019) have been developed for English. For many smaller languages the only available options are multilingual models, which

can reach impressive performance on downstream tasks (Conneau et al., 2020). These are not without their flaws though; when compared to training on a sufficiently large monolingual corpus in a given language, multilingual models can lead to less than optimal performance, as demonstrated in the case of Finnish (Virtanen et al., 2019). Since an evaluation of Transformer language models for Icelandic is yet to be completed, it has remained unclear to what extent this holds for Icelandic.

The data used to train language models is usually sourced from large collections of books (e.g. (Zhu et al., 2015)) and online texts, where the choice and quality of training data can potentially have a large effect on downstream task performance. While curated corpora may not be readily available for a language, it might still be relatively well represented online, in the form of web texts, which can be sourced by automatic means. This raises the question of whether language models trained on curated corpora offer better performance in downstream tasks than those trained predominantly on data sourced from the web.

In this paper, we show how a Transformer model, IceBERT, can be trained for Icelandic with relatively modest language resources to reach state-of-the-art performance across a variety of tasks. We train multiple models on monolingual corpora from different sources: a curated corpus (Icelandic Gigaword Corpus, IGC (Steingrímsson et al., 2018)) and a corpus of text collected efficiently from Common Crawl[2].

We train separate models on each of these two sources,

---

* Equal contribution.

[1]*Low-resource* is not a precise term, but a language can be considered to be low-resource if few online resources exist for it (Cieri et al., 2016).

[2]`commoncrawl.org`

and compare results, to demonstrate the feasibility of our approach for other languages with similar resource availability. To our surprise, models trained on texts extracted from the web achieved similar performance to models trained on a curated corpus. We also evaluate the performance of a multilingual model (XLMR-base, Conneau et al. (2020)), which shows good results for some tasks but is insufficient for others. Finally, we use the existing multilingual model as a warm start and continue pre-training on Icelandic text, reaching state-of-the-art results in downstream tasks such as NER.

While large corpora of multilingual text exist, such as the multilingual Colossal Clean Crawled Corpus (mC4) (Xue et al., 2021) that is sourced from the Common Crawl, they have not been officially released in a way such that text in smaller languages can be easily extracted. However, the mC4 dataset has been made available by a third party[3] which we use for our experiments. Additionally, we demonstrate how to directly extract Icelandic text from the Common Crawl in a novel way, explain how it can be done for other languages, and highlight the importance of clean data.

Regarding applicability of our approach to other languages, we note that in mC4, there are 107 labelled languages, with almost half of the 6.7 billion documents being in English. The average number of documents in languages other than English is 33.4 million documents per language, and the median value is 2 million documents, which happens to be the approximate number for Icelandic (Xue et al., 2021). These numbers show that our approach of extracting training data should be well within reach of at least half of the languages in Common Crawl, and possibly applicable for the 46 languages containing between 500 thousand and 10 million documents.

The key contributions of our work are summarized below.

(a) Several Icelandic language models, including IceBERT,[4] trained on a monolingual corpus with 2.7B tokens.

(b) Adaptations of IceBERT with state-of-the-art results for part-of-speech tagging (PoS), named entity recognition (NER), constituency parsing and grammatical error detection (GED).

(c) The Icelandic Common Crawl Corpus (IC3)[5], a cleaned and deduplicated corpus extracted by targeting the `.is` top level domain.

(d) The Icelandic WinoGrande dataset (IWG)[6], a new and challenging benchmark for commonsense reasoning and natural language understanding.

---

2. Related work

The original BERT model has, since its publishing, spawned a whole family of BERT-like models. One of the main reasons for their popularity is their potential for transfer learning, i.e. the possibility to adapt them and obtain impressive performance on benchmarks and tasks that they were not originally trained for.

Multilingual versions of BERT exist that are trained on text in multiple languages, such as mBERT, which is trained on Wikipedia in over a hundred languages, including Icelandic. Since the release of BERT, other large pre-trained models such as mBART (Liu et al., 2020) and XLMR (Conneau and Lample, 2019; Conneau et al., 2020) have been trained that include Icelandic and other lower-resource languages. In addition, mT5 (Xue et al., 2021), a sequence-to-sequence model, is trained on the entire mC4 corpus.

Multilingual models are often the only option for low-resource languages, which do not have direct access to sufficient language data or computational resources to create monolingual transformer-based language models. Such multilingual models have been shown to have useful properties, including zero-shot crosslingual transfer. That is, fine-tuning these models on a downstream task in one language can translate to improved performance in other languages without explicit crosslingual signals (Pires et al., 2019; Wu and Dredze, 2019; Karthikeyan et al., 2019).

Despite the impressive results for multilingual models, they might not be the right choice where output accuracy is critical. For some languages, it may be better to pre-train a model on a monolingual corpus and adapt it for downstream tasks rather than to adapt a model trained on a multilingual corpus, as in the case of Finnish (Virtanen et al., 2019). It has also been shown that the crosslingual capabilities of mBERT only apply to high-resource languages (Wu and Dredze, 2020). Furthermore, benchmarks of crosslingual transfer show a sizable gap in the performance of crosslingually transferred models when compared to monolingually trained ones (Hu et al., 2020). These results highlight the still basic need for more training data in the case of medium- and low-resource languages.

As a result, work has been ongoing in establishing baselines and mapping the performance of monolingual models. Some of that work on high-resource languages is summarized by Scheible et al. (2020). We highlight examples of published monolingual models for medium and low-resource languages along with English in Table 1 with an emphasis on languages with resources similar to Icelandic in mC4. Generally, the model building approach is similar, although we note that in one case a multilingual model was used as a warm start (Ralethe, 2020).

We would also like to point out that for several languages of similar size to Icelandic (2.6B tokens) in mC4, there is no public monolingual BERT model available. These include Maltese (5.2B tokens),

Language	mC4 tokens (B)	Native speakers	Reference
English	2,733	380M	(Devlin et al., 2019)
Icelandic	2.6	350k	(This paper)
Galician	2.4	2.4M	(Vilares et al., 2021)
Urdu (Roman)	2.4	70M	(Khalid et al., 2021)
Filipino	2.1	23.8M	(Cruz and Cheng, 2020)
Afrikaans	1.7	7.2M	(Ralethe, 2020)
Basque	1.4	900k	(Agerri et al., 2020)
Telugu	1.3	83M	(Marreddy et al., 2021)
Latin	1.3	0	(Bamman and Burns, 2020)
Swahili	1.0	18M	(Bhattacharjee et al., 2021)

Table 1: Language models trained on a monolingual corpus. We highlight some languages that have a similar number of tokens to Icelandic in mC4. The number of speakers denotes the number of native speakers (L1) according to the Wikipedia page for each language.

Kazakh (3.1B tokens), Georgian (2.5B tokens), Belarusian (2B tokens), Tajik (1.4B tokens), Kyrgyz (1B tokens), Somali (1.4B tokens), Sindhi (1.6B tokens), Armenian (2.4B tokens), and Luxembourgish (1B tokens). For others languages such as Macedonian (1.8B tokens), Malayalam (1.8B tokens), Mongolian (2.7B tokens), and Kannada (1.1B tokens) models exist online but to our best knowledge no publications exist that thoroughly document their performance on basic downstream tasks, such as PoS tagging and NER.

Another line of research has focused on how to make multilingual models more effective for low-resource languages. It has been shown that training on a larger corpus, such as filtered Common Crawl data, leads to significant improvements in downstream tasks, but increasing the number of languages beyond a certain point has a diluting effect that reduces overall performance (Conneau et al., 2020). Others have shown that vocabulary extension of a multilingual model with continued pre-training on a monolingual corpus leads to improved performance and shorter training times than when starting from scratch (Wang et al., 2020).

The results on multilingual models indicate that language model performance is related to the amount of training data for the given language (Xue et al., 2021), and studies on corpus quality indicate that the results are strongly related to the number of high quality sentences (Kreutzer et al., 2022). Kreutzer et al. (2022) have emphasized the importance of evaluating and auditing the corpora that are publicly available, since data in low-resource languages from multilingual datasets can be of low quality. They further emphasize the importance of developing high-quality evaluation datasets, since low-quality benchmarks might exaggerate model performance, making NLP for low-resource languages look further developed than it actually is.

2.1. NLP for Icelandic

Icelandic is a language from the West Germanic language family, with a rich morphology, where nouns, adjectives and verbs are highly inflected, and com-

pounding is used actively to construct new words. The status of language data and resources for Icelandic is steadily improving, providing us with various datasets for evaluating our models, and benchmarks to measure against.

A good deal of work has been done on NLP for Icelandic that concerns these benchmarks. PoS tagging is implemented using a rule-based approach in the IceNLP toolkit (Loftsson and Rögnvaldsson, 2007), and using a Bi-LSTM model in the ABLTagger (Steingrímsson et al., 2019). Constituency parsing has been implemented using a hand-crafted context-free grammar in the Greynir package (Þorsteinsson et al., 2019), using finite-state transducers in IceParser (Loftsson and Rögnvaldsson, 2007), and using an mBERT model in (Arnardóttir and Ingason, 2020). NER for Icelandic has been implemented using a Bi-LSTM model and an ensemble tagger (Ingólfsdóttir et al., 2020).

3. Training data

The Icelandic datasets used for pre-training our models are listed in Table 2. They were randomly split at the document level into train (80%), validation (5%) and holdout (15%) splits, and then tokenized (Þorsteinsson et al., 2020). We do not use the large holdout in this paper but reserve it for future experiments. We also do experiments on the Icelandic subset of the mC4 dataset (not shown in the table, see Section 3.2), a dataset that is, in similar fashion to IC3, extracted from the Common Crawl but via a different method.

The IGC (Steingrímsson et al., 2018) is the most extensive collection of curated Icelandic text available. The IGC is mostly made up of news, legal documents and other copy-edited content and might, therefore, not accurately reflect the distribution of text from online sources. To supplement this dataset, several other sources were collected for pretraining the language models, as listed in Table 2. At the time of training the model, large collections of Icelandic literature were not available through legal means, but the recently updated IGC now includes some literary texts (Barkar-

Dataset	Size	Tokens
IGC (editorial text)	8.2GB	1,388M
IC3 (cleaned webcrawl)	4.9GB	824M
Student theses	2.2GB	367M
Greynir News articles	456MB	76M
Medical library	33MB	5.2M
Open Icelandic e-books	14MB	2.6M
Icelandic Sagas	9MB	1.7M
Total	15.8GB	2,664M

Table 2: Icelandic texts used for training.

son et al., 2021b), which will be incorporated in future models. Social media and internet forum texts have now also been added to the updated IGC (Barkarson et al., 2021a).

The IC3, our corpus of scraped and cleaned web texts (see next section for details), contains large amounts of text of many domains, topics, and styles at varying degrees of polish, and thus serves well as a complement to the IGC. In addition to the already mentioned data, academic texts found in student theses[7] and data from the medical library of the University Hospital of Iceland[8] were collected. The academic texts were passed through a filter reminiscent of the one used for the IC3 described in the next section, after an initial PDF text-extraction step. We also use texts scraped from Icelandic online news sites by the Greynir NLP engine[9].

3.1. The Icelandic Common Crawl Corpus

The Common Crawl Foundation is a non-profit organization that scrapes large semi-random subsets of the internet regularly and hosts timestamped and compressed dumps of the web online[10]. Each dump contains billions of web pages occupying hundreds of terabytes. Parsing these files directly requires storage and computing power not directly available to most and can come at a significant financial cost. The foundation also hosts indices of URIs and their locations within the large zipped dump files. While these indices are also large, their processing is feasible with a few terabytes of storage.

3.1.1. Extracting Icelandic Common Crawl data

The Common Crawl indices, which contain URI and byte offsets within the compressed dumps, are used to reduce the search space when looking for Icelandic texts. The Common Crawl Index Server has a public API[11] where URIs can be queried based on attributes such as date, MIME-type and substring. Using the API eliminates the need to fetch the massive index files.

---

[7]skemman.is
[8]www.hirsla.lsh.is
[9]greynir.is
[10]commoncrawl.org/the-data/get-started/
[11]index.commoncrawl.org

To extract Icelandic, the .is pattern is targeted to match the Icelandic top level domain (TLD), resulting in 63.5 million retrieved pages with URIs and byte locations within the compressed Common Crawl dumps. The computational efficiency of our method can be attributed to these steps. Given the predominant use of the .is TLD for Icelandic web content, we assume that other TLDs have a much lower proportion of Icelandic content. That said, a nontrivial amount of text in Icelandic is still likely to be found outside the .is domain and could be extracted by, e.g., parsing the whole Common Crawl, albeit at a much higher computational cost.

By targeting only the byte-offsets corresponding to the Icelandic TLD we extract candidate websites that have a high proportion of Icelandic content. In total, the compressed content is 687GiB on disk. All dumps since the start of the Common Crawl in 2008 until March 2020 were included.

Plain text was extracted from the collected WARC (Web Archive format) files using jusText (Pomikálek, 2011)[12] to remove boilerplate content and HTML tags.

3.1.2. Processing Common Crawl

Once plain text had been extracted from the WARC files, Icelandic text was taken aside and duplicates removed. Since the .is TLD contains text in numerous languages, we use a fastText (Bojanowski et al., 2017) model for extracting Icelandic text. Since the web is abundant with duplicate or near duplicate content, the data is first deduplicated at the document level and then at the inter-sentence level by sliding a three-line window over the text. If any three consecutive lines have appeared together previously, they are discarded. This latter step removes a fair amount of unwanted content, such as cookie notifications and thumbnail text. A summary of the filtering steps taken is shown in Table 3.

Filtering step	Size	%
.is TLD	687GB	100%
IS lang. filter and boilerpl. rem.	29GB	4.2%
Dedup. document	8.6GB	1.3%
Dedup. window	4.9GB	0.71%

Table 3: Filtering steps and retained data for IC3.

3.1.3. Comparison between IGC and IC3

The two corpora, IC3 and IGC, are significantly different at the level of individual words. There are 1,155k unique tokens in the IC3 and 1,434k unique tokens in IGC, of which only 818k are shared.[13] Almost one-third of the unique tokens (337k) in the IC3 are not

---

[12]We use the implementation at https://github.com/miso-belica/jusText.
[13]If tokens with a count below 5 are not excluded there are 6.5M unique tokens in IGC and 6M unique tokens in IC3.

present in IGC, and almost half of the IGC tokens (616k) are not present in IC3.

3.2. The Icelandic part of mC4

The Icelandic part of mC4 (mC4-is) contains 2.6B tokens (~8GB on disk). The data was not filtered in the same way as the IC3 which is reflected by the masked token perplexity results shown in Table 4; we hypothesize that further processing would be necessary to make use of it. By eyeing a random subset of the data, we see that a fair amount is badly machine-translated and some segments contain a lot of noise that is non-alphanumeric or otherwise not fluid text. If further processed, the results might very well be similar to that of IC3, but this analysis is left as future work.

4. Training language models

We train four different BERT-base models following the hyperparameter setup in RoBERTa (Liu et al., 2019). We use 48 Nvidia 32GB V100 GPUs for approximately two days or 225k updates, with a batch size of ~955k tokens (2k sequences). We also train a single model using the RoBERTa-large architecture. This model became unstable in training after 37.5k steps with a batch size of ~4M tokens (8k sentences) and we did not make attempts to further improve it, but we fine-tune it in our experiments for comparison with the base models. All models use the same BPE-vocabulary, containing 50k tokens, constructed in the same way as the RoBERTa vocabulary. We train the models using four different data settings: all of the data available except mC4-is (IceBERT and IceBERT-large); the IC3 dataset (IceBERT-IC3), the IGC dataset (IceBERT-IGC); and mC4-is (IceBERT-mC4-is).

Furthermore, we evaluate performance using the multilingual XLMR-base (Conneau et al., 2020) model as-is. We also experiment with continued pre-training on the IC3 corpus. Two models are trained: One for 100k steps with a batch size of 40k tokens and the other for 225k steps with a batch size of 80k tokens. The first model took one GPU-day in training and the other seven GPU-days. We do this to show what performance can be gained from leveraging a publicly available multilingual model while minimizing the computation cost as the seven day model uses about 8% of the GPU hours used for training of the IceBERT-base models.

5. Results

We fine-tune and adapt the different IceBERT models for several classification and parsing tasks with state-of-the-art results after fine-tuning. Where F-scores are reported they are macro-averaged.

For PoS labelling and constituency parsing we extend the `fairseq` library; the resulting package `greynirseq` has been made available[14]. We use the

Dataset	All*	IC3	IGC	mC4-is
IGC	3.64	4.49	3.51	16.50
IC3	4.40	4.15	5.67	17.84
Student theses	4.12	4.95	5.47	15.40
Medical library	5.26	5.87	7.26	20.49
Greynir News	3.71	4.43	4.05	14.11
Icelandic Sagas	7.53	7.09	12.49	52.00
Icelandic e-books	9.04	8.75	10.37	36.48

Table 4: Masked token perplexity over validation sets using models trained on different Icelandic datasets and combinations thereof. The All* model refers to IceBERT trained on all data except mC4-is.

implementation in `fairseq` to evaluate performance on the Icelandic WinoGrande dataset.

For named entity recognition and grammatical error detection, we use the `transformers` library from Hugging Face (Wolf et al., 2020). For this purpose, we convert IceBERT to work with the library[15].

5.1. Part of Speech

We fine-tune our models for PoS tagging using `greynirseq` on the MIM-GOLD (Barkarson et al., 2020) dataset using ten-fold cross validation. The best performing models reach an accuracy of 98.4%. We exclude the *x* (not analyzed due to e.g. incorrect spelling) and *e* (foreign) labels. In comparison, Steingrímsson et al. (2019) achieve 94.04% accuracy. The results are shown in Table 5.

In contrast to prior work on Icelandic PoS tagging, which universally approaches this as a multi-class classification task, we use a multi-label multi-class approach where we predict grammatical categories (gender, tense, etc.) independently instead of all together in one label. We adopt this approach to address a significant label scarcity problem in the training set and to allow for better generalization. See appendix 7 for a more comprehensive description.

We train the PoS models with a batch size of 32 sentences for 5 epochs. Peak learning rate is 5-e5 with approximately 0.2 epochs for warmup and a linear decay to zero. For the randomly initialized (no pretraining) model we do an additional longer run on the data since the model was clearly nowhere near convergence after 5 epochs, this was only done on a single split of data due to time constraints.

All of the Icelandic models (except the one trained on mC4-is) show similar results of ~98.3% accuracy. An informal review of the errors that the best models make leads us to believe that this performance is about as good as it gets with this dataset and model architecture. The majority of errors can be classified as either being problems with the reference data or due to inherently ambiguous sentences. Problems with the reference data

---

[14]See `github.com/mideind/greynirseq` and the package `greynirseq` on PyPI

[15]The resulting model is available for use at `huggingface.co/mideind/IceBERT`.

Model	Accuracy
IceTagger (2007)	91.5
ABLTagger (2019)	95.15
IceBERT	98.33 ± 0.05
IceBERT-large	98.35 ± 0.06
IceBERT-IGC	98.27 ± 0.05
IceBERT-IC3	98.30 ± 0.05
IceBERT-mC4-is	97.62 ± 0.10
XLMR-base	96.70 ± 0.15
XLMR-base-IC3-1d	97.20 ± 0.10
XLMR-base-IC3-7d	98.20 ± 0.07
No pretraining (5 epochs)	74.96 ± 0.54
No pretraining (50 epochs)	90.27

Table 5: Comparison of PoS-tagging performance for the models considered, including randomly initialized models.

are either mislabeled examples or inconsistently applied rules, especially around proper nouns. Ambiguous sentences are mostly due to pronouns whose gender is unknowable without longer context. Most of the remaining errors are very difficult examples that require extensive world knowledge or complex co-reference resolution.

5.2. Named Entity Recognition

When fine-tuning IceBERT for named entity recognition (NER) on the Icelandic NER dataset (Ingólfsdóttir et al., 2020), it reaches state-of-the-art performance, showing a considerable improvement over the prior result of 85.79 macro F1-score achieved by Ingólfsdóttir et al. (2020). In fine-tuning we use a batch size of 16 sentences, peak learning rate of "2e-5 and chose the highest performing model on the validation set as measured across 10 epochs. The results over the test set for the different models each averaged over five seeds are shown in table 6.

The results are similar for all monolingual models and show that a lot of data or curated editorial corpora in pre-training are not necessary to achieve competitive NER performance. We would like to highlight that the XLMR-base multilingual model trained for 7-days on IC3 performs best on this task.

5.3. Constituency parsing

We implement a simplified version of the CKY-style chart parser described by Kitaev et al. (2019)[16]. We did not implement position-factored attention nor incorporate any extra word features, such as PoS or character information, since our goal is primarily to measure the knowledge captured by the model. We leave such experiments for future work.

The dataset we use is GreynirCorpus (Þorsteinsson et al., 2021), a constituency annotated version of

---

[16]Implemented in the `greynirseq` repository.

the aforementioned Greynir News dataset, whose test and validation sets are human-annotated. Its annotation scheme comes from the Greynir rule-based parser (Þorsteinsson et al., 2019) and shares many similarities with the Penn Treebank-derived (Marcus et al., 1993) schemas and their corresponding annotation guidelines.

A generalized version of the GreynirCorpus test set was created for a fairer comparison with previous parsers for Icelandic, namely the Greynir parser, IceParser (Loftsson and Rögnvaldsson, 2007) — a shallow parser for Icelandic, and a variant of the Berkeley Neural Parser (Arnardóttir and Ingason, 2020) which comprises a multilingual BERT fine-tuned on the Icelandic Parsed Historical Corpus (IcePaHC) (Rögnvaldsson et al., 2012).

We split the validation subset of the GreynirCorpus into ad-hoc train and validation splits and train on the respective portion. Results from testing on the generalized benchmark mentioned above are shown in Table 7 as measured using evaluation of bracketing (EVALB), which is a score based on matching brackets (Sekine and Collins, 2008). Somewhat surprisingly, the large model does not show best performance on the task, we believe that this would change if the model is trained to convergence or better hyperparameter tuning. We note that differences between models are only slight as all the models are within 2 percentage points from each other. For extra comparison, a randomly initialized model did not surpass 70 F1-score.

Note that in earlier published evaluations of Icelandic constituency parsing an older version of the test set was used. It was also smaller in the sense that sentences with with grammatical errors were excluded. Additionally in the case of the Berkeley neural parser, all sentences with multiword tokens such as "í gær" (yesterday) were removed, effectively halving the number of test sentences. On that test set the Greynir parser had an F1 score of 81.21 and a fine-tuned IceBERT-IGC model had an F1 score of 90.67, the Berkeley neural parser variant had an F1 score of 87.07 (this differs from the result in Table 7 as a different test set was used for the comparison). The IceParser is a shallow parser and thus cannot be evaluated using EVALB.

5.4. Grammatical error detection

We fine-tune IceBERT for grammatical error detection (GED). These are the first machine learning models trained for GED in Icelandic, and make use of the Icelandic Error Corpus (IceEC) (Ingason et al., 2021) which contains 58k labeled sentences.

We train models for both binary and multi-class token-level classification. For the multi-class GED classifier, we exclude ambiguous sentences from the IceEC where a single token has multiple error labels; this removes 3.5k sentences out of the 23k sentences with error labels. The problem could be solved as a multi-label one, but we leave that approach for future work.

Model	F1	Prec.	Rec.	Acc.
Bi-LSTM (Ingólfsdóttir et al., 2020)	85.79	-	-	-
IceBERT	91.43 ± 0.23	91.60 ± 0.13	91.26 ± 0.36	98.66 ± 0.03
IceBERT-large	91.20 ± 0.36	90.79 ± 0.84	91.61 ± 0.41	98.58 ± 0.08
IceBERT-IGC	91.10 ± 0.25	91.15 ± 0.38	91.06 ± 0.20	98.59 ± 0.04
IceBERT-IC3	91.29 ± 0.16	91.24 ± 0.24	91.35 ± 0.27	98.62 ± 0.02
IceBERT-mC4-is	89.57 ± 0.28	89.27 ± 0.44	89.87 ± 0.28	98.40 ± 0.06
XLMR-base	88.95 ± 0.60	88.66 ± 0.78	89.25 ± 0.61	98.41 ± 0.08
XLMR-base-IC3-1d	89.58 ± 0.30	89.84 ± 0.46	89.33 ± 0.26	98.39 ± 0.04
XLMR-base-IC3-7d	92.52 ± 0.40	92.31 ± 0.49	92.74 ± 0.41	98.83 ± 0.05

Table 6: NER performance for models trained on different datasets, standard deviation over five seeds shown.

Model	F1	Prec.	Rec.
IceBERT	90.02 ± 0.12	87.93 ± 0.16	92.20 ± 0.07
IceBERT-large	89.79 ± 0.13	87.71 ± 0.28	91.98 ± 0.14
IceBERT-IGC	89.66 ± 0.12	87.20 ± 0.13	92.25 ± 0.15
IceBERT-IC3	89.37 ± 0.14	86.73 ± 0.32	92.18 ± 0.17
IceBERT-mC4-is	88.60 ± 0.16	86.08 ± 0.13	91.27 ± 0.40
XLMR-base	88.16 ± 0.27	86.16 ± 0.33	90.26 ± 0.24
XLMR-base-IC3-1d	88.67 ± 0.16	86.74 ± 0.10	90.75 ± 0.24
XLMR-base-IC3-7d	89.01 ± 0.09	86.95 ± 0.12	91.16 ± 0.09

Table 7: EVALB performance on the generalized form of the GreynirCorpus test set, mean and standard deviation over five seeds shown.

While the dataset is fine-grained and contains a variety of objective and subjective labels (e.g. stylistic), in this study we limit our evaluation to the five high-level categories *coherence, grammar, orthography, style and vocabulary*.

For fine-tuning, we use a batch size of 16, learning rate of 2e-5 and train five times for five epochs with different seeds. The results for binary classification are shown in Table 8 and by category for the top 5 models in the multi-class task in Table 9. The three lowest performing models are not shown in Table 9 due to lack of space. For the sake of completeness, IceBERT-mC4-is had a total accuracy of 51.69 ± 2.60, XLMR had a total accuracy of 56.26 ± 4.22, and XLMR-IC3-1d had a total accuracy of 41.24 ± 0.60.

Based on the results in Tables 8 and 9 it is clear that out of the IceBERT base-models, the model trained on the IGC dataset containing editorial text is best suited for fine-tuning for GED. IceBERT-large is the clear winner with an F1 score of 89.12 ± 1.31 in binary token classification and 79.16 ± 9.39 in token classification. The 1-day XLMR model does not do as well as the IceBERT models and the mC4-is model is lagging behind, further highlighting that models trained on the dataset might benefit from further cleanup. Interestingly, we see that the 7-day XLMR multilingual model outperforms the other models besides IceBERT-large. The improvement of the the 7-day XLMR model in token classification when compared with IceBERT-base

models is mainly due to better results in the categories grammar and style as can be seen in Table 9.

5.5. Icelandic WinoGrande

The WinoGrande dataset (Sakaguchi et al., 2020), used for evaluating commonsense reasoning capabilities of neural language models, is inspired by the original WinoGrad dataset (Levesque et al., 2012), but its problems are designed to minimize biases which the models may rely on when solving them. The dataset consists of sentences that include two nouns and an ambiguous pronoun which grammatically can refer to either of those noun phrases. The task is to decide which noun makes more semantic sense, given the information in the sentence.

We systematically go through the WinoGrande test set (1767 examples) and manually translate and adapt sentences to work in Icelandic. While the English WinoGrande problems are not always constructed as pairs, in our adaptation, we create sentence pairs where it is feasible. We also found some of the examples to be specific to culture, subjective, or otherwise inapplicable for translation. Those examples were either adjusted or skipped. The result is a dataset of 1095 examples. The size of the Icelandic dataset is closest in size to the small variant of the English dataset (640 examples).

The results over five-fold cross-validation can be seen in Table 10. The large model outperforms the other variants while models trained on IGC and/or IC3 out-

Model	F1	Prec.	Rec.	Acc.
IceBERT	70.11 ± 0.91	92.30 ± 0.17	56.53 ± 1.12	96.99 ± 0.07
IceBERT-large	89.12 ± 1.31	93.55 ± 0.67	85.10 ± 1.87	98.70 ± 0.15
IceBERT-IGC	71.33 ± 1.65	92.14 ± 0.42	58.22 ± 2.08	97.08 ± 0.13
IceBERT-IC3	70.95 ± 1.10	91.97 ± 0.62	57.77 ± 1.41	97.05 ± 0.09
IceBERT-mC4-is	57.72 ± 1.07	90.98 ± 0.58	42.28 ± 1.20	96.13 ± 0.06
XLMR-base	64.52 ± 1.82	88.11 ± 0.72	50.93 ± 2.21	96.75 ± 0.12
XLMR-base-IC3-1d	62.73 ± 3.05	86.62 ± 1.55	49.22 ± 3.38	96.61 ± 0.21
XLMR-base-IC3-7d	75.87 ± 1.02	91.64 ± 0.23	64.74 ± 1.43	97.61 ± 0.08

Table 8: Binary token classification performance measured using the Icelandic Error Corpus evaluation dataset, standard deviation over five seeds shown.

Category	IB-base	IB-large	IB-IGC	IB-IC3	XLMR-IC3-7d
Coherence	5.90 ± 3.15	41.58 ± 21.83	6.89 ± 3.68	4.94 ± 3.23	2.40 ± 3.00
Grammar	55.05 ± 2.64	72.09 ± 14.25	54.71 ± 4.77	50.71 ± 4.93	62.83 ± 3.69
Orthography	80.39 ± 2.12	89.66 ± 4.88	81.00 ± 2.30	79.79 ± 2.29	84.47 ± 2.19
Style	36.07 ± 10.42	67.39 ± 16.30	37.19 ± 9.66	35.50 ± 9.66	47.44 ± 9.02
Vocabulary	18.47 ± 2.74	50.30 ± 21.45	17.94 ± 4.48	16.85 ± 4.45	17.33 ± 4.56
All	62.83 ± 3.75	79.16 ± 9.39	63.45 ± 3.96	61.84 ± 4.06	69.42 ± 3.58

Table 9: Token classification F1-scores measured using the Icelandic Error Corpus evaluation dataset for the top-5 highest scoring models. IB is shorthand notation for IceBERT.

Model	Accuracy
IceBERT	54.6 ± 2.1 %
IceBERT-large	57.1 ± 3.7 %
IceBERT-IGC	53.8 ± 2.3 %
IceBERT-IC3	53.8 ± 2.4 %
IceBERT-mC4-is	51.4 ± 2.1 %
XLMR-base	52.4 ± 1.7 %
XLMR-base-IC3-1d	51.2 ± 2.0 %
XLMR-base-IC3-7d	53.4 ± 1.9 %

Table 10: Accuracy on the Icelandic WinoGrande dataset. Results are averaged over five-fold cross-validation and standard deviation is reported.

which show performance that is competitive with those trained on curated/editorial corpora. We stress that proper filtering and cleanup of crawled data is necessary, as demonstrated in the difference between training models on IC3 and mC4-is.

For some downstream tasks, we observe that multilingual language models (Conneau et al., 2020) are getting quite close to the performance of models trained on a monolingual corpus but for other tasks we still observe a significant difference. Interestingly, using such models as a warm start for some tasks can even lead to state-of-the-art performance.

We conclude that by using text extracted from the Common Crawl corpus and multilingual models as a warm start, well-performing language models are becoming more feasible to build for low to medium-resource languages. We still hold that curated corpora are beneficial in certain applications, but the added value for downstream tasks is small if a sufficiently large crawled corpus is available.

perform the model trained on mC4-is. The XLMR-base and XLMR-base-IC3-1d perform similar to the model trained on mC4-is but the XLMR-base-IC3-7d performs similar to the model trained only on IC3 or IGC.

6. Conclusion

We have successfully trained baseline neural language models for Icelandic that perform well on existing benchmarks, in particular NER, PoS and constituency parsing. We also present the Icelandic WinoGrande dataset and show that it is challenging for the models we evaluate.

Furthermore, to our surprise, we show that extracting data from online sources is sufficient to train models

7. Acknowledgements

Appendix: Part of Speech tagging

We use the MIM-GOLD dataset (Barkarson et al., 2020) to train a PoS tagger. If one approaches this task as a multiclass problem (each word gets exactly one label) there is a significant label scarcity problem. There are approximately 600 legal labels in the MIM-GOLD schema, but only 559 of them appear in the training set, and 43 of them appear less than 10 times. The 10th, 25th, 50th and 75th percentile tags occur 15, 52, 208 and 728 times, respectively.

To address this problem we decompose the labels. Instead of considering a label to be a single unit, e.g. *noun-masculine-singular-nominative-article*, we consider it to be composed of several categories, e.g. *lexical class*, *gender*, *number*, *case* and *article-clitic*, each of which can have several values. We also observe that some categories are shared between lexical class, e.g. nouns and adjectives both have gender, number and case and typically share values when they co-refer. Our model therefore outputs for each word a lexical class and a value for every grammatical category or morphological feature, but we ignore those that are not applicable to the lexical class, e.g. for nouns we mask out loss for the *tense* category during training and output no *tense* label during inference.

Since the number of labels within each category is small (the largest category has 6 possible labels), each label has been seen many times during training, even though some combinations of labels never occur in the training set (such as *verb past subjunctive 2person plural middle-voice*). This allows the model to generalize and predict these unseen combinations, some of which actually occur in the test set.

8. Bibliographical References

Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., and Agirre, E. (2020). Give your text representation models some love: the case for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788.

Arnardóttir, Þ. and Ingason, A. K. (2020). A Neural Parsing Pipeline for Icelandic Using the Berkeley Neural Parser. In Costanza Navarretta et al., editors, *Proceedings of CLARIN Annual Conference 2020*, pages 48–51.

Bamman, D. and Burns, P. J. (2020). Latin BERT: A Contextual Language Model for Classical Philology. *arXiv:2009.10053 [cs]*, September. 00007 arXiv: 2009.10053.

Bhattacharjee, A., Hasan, T., Samin, K., Islam, M. S., Rahman, M. S., Iqbal, A., and Shahriyar, R. (2021). BanglaBERT: Combating Embedding Barrier in Multilingual Models for Low-Resource Language Understanding. *arXiv:2101.00204 [cs]*, August. 00005 arXiv: 2101.00204.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., and et al. (2021). On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Cieri, C., Maxwell, M., Strassel, S., and Tracey, J. (2016). Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Conneau, A. and Lample, G. (2019). Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. 00847.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Cruz, J. C. B. and Cheng, C. (2020). Establishing Baselines for Text Classification in Low-Resource Languages. *arXiv:2005.02068 [cs]*, May. 00010 arXiv: 2005.02068.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421. PMLR, November. 00263 ISSN: 2640-3498.

Ingólfsdóttir, S. L., Guðjónsson, Á. A., and Loftsson, H. (2020). Named Entity Recognition for Icelandic: Annotated Corpus and Models. In Luis

Espinosa-Anke, et al., editors, *Statistical Language and Speech Processing*, Lecture Notes in Computer Science, pages 46–57, Cham. Springer International Publishing. 00002.

Karthikeyan, K., Wang, Z., Mayhew, S., and Roth, D. (2019). Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *International Conference on Learning Representations*.

Khalid, U., Beg, M. O., and Arshad, M. U. (2021). RUBERT: A Bilingual Roman Urdu BERT Using Cross Lingual Transfer Learning. *arXiv:2102.11278 [cs]*, February. 00000 arXiv: 2102.11278.

Kitaev, N., Cao, S., and Klein, D. (2019). Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy, July. Association for Computational Linguistics.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., Orife, I., Ogueji, K., Rubungo, A. N., Nguyen, T. Q., Müller, M., Müller, A., Muhammad, S. H., Muhammad, N., Mnyakeni, A., Mirzakhalov, J., Matangira, T., Leong, C., Lawson, N., Kudugunta, S., Jernite, Y., Jenny, M., Firat, O., Dossou, B. F. P., Dlamini, S., de Silva, N., Çabuk Ballı, S., Biderman, S., Battisti, A., Baruwa, A., Bapna, A., Baljekar, P., Azime, I. A., Awokoya, A., Ataman, D., Ahia, O., Ahia, O., Agrawal, S., and Adeyemi, M. (2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, January.

Kristinsson, A. P. (2018). National language policy and planning in iceland–aims and institutional activities. In *National Language Institutions and National Languages. Contributions to the EFNIL Conference 2017*, pages 243–249.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neu-ral machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Loftsson, H. and Rögnvaldsson, E. (2007). Icenlp: a natural language processing toolkit for icelandic. In *INTERSPEECH*, pages 1533–1536.

Loftsson, H. and Rögnvaldsson, E. (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 128–135, Tartu, Estonia, May. University of Tartu, Estonia. 00033.

Marreddy, M., Oota, S. R., Vakada, L. S., Chinni, V. C., and Mamidi, R. (2021). Clickbait Detection in Telugu: Overcoming NLP Challenges in Resource-Poor Languages using Benchmarked Techniques. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July. 00000 ISSN: 2161-4407.

Nikulásdóttir, A., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. (2020). Language technology programme for Icelandic 2019-2023. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France, May. European Language Resources Association.

Þorsteinsson, V., Óladóttir, H., and Loftsson, H. (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, September. INCOMA Ltd.

Þorsteinsson, V., Óladóttir, H., and Þórðarson, S. (2020). Tokenizer for icelandic text. CLARIN-IS.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Pomikálek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora [online]*. Doctoral theses, dissertations, Masaryk University, Faculty of Informatics, Brno.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Ralethe, S. (2020). Adaptation of Deep Bidirectional Transformers for Afrikaans Language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2475–2478, Marseille, France, May. European Language Resources Association. 00002.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2012). The Icelandic parsed historical corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V.,

and Boeker, M. (2020). GottBERT: a pure German Language Model. *arXiv:2012.02110 [cs]*, December. 00009 arXiv: 2012.02110.

Sekine, S. and Collins, M. (2008). Evalb.

Steingrímsson, S., Kárason, Ö., and Loftsson, H. (2019). Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Bulgaria, September.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vilares, D., Garcia, M., and Gómez-Rodrıguez, C. (2021). Bertinho: Galician BERT Representations. *Procesamiento del Lenguaje Natural*, 66:13–26.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv:1912.07076 [cs]*, December. arXiv: 1912.07076.

Wang, Z., Karthikeyan, K., Mayhew, S., and Roth, D. (2020). Extending Multilingual BERT to Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Wu, S. and Dredze, M. (2019). Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China, November. Association for Computational Linguistics. 00329.

Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

9. Language Resource References

Barkarson, S., Sigurðsson, E. F., Rögnvaldsson, E., Hafsteinsdóttir, H., Loftsson, H., Steingrímsson, S., and Andrésdóttir, Þ. D. (2020). *MIM-GOLD 20.05*.

Barkarson, S., Steingrímsson, S., and Daníelsson, H. (2021a). *IGC-Social 21.10*.

Barkarson, S., Steingrímsson, S., Hafsteinsdóttir, H., and Ingimundarson, F. (2021b). *IGC-Books 21.10*.

Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., and Xu, X. (2021). *The Icelandic Error Corpus (IceEC), version 1.1*.

Ingólfsdóttir, S. L., Guðjónsson, Á. A., and Loftsson, H. (2020). MIM-GOLD-NER – named entity recognition corpus (20.06). CLARIN-IS.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). *Building a Large Annotated Corpus of English: The Penn Treebank*. MIT Press.

Þorsteinsson, V., Óladóttir, H., Þórðarson, S., Símonarson, H. B., and Ásgeirsdóttir, K. (2021). *Greynir-Corpus (2021-06-23)*.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2020). *WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale*.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. (2018). *Risamálheild: A Very Large Icelandic Text Corpus*. European Language Resources Association (ELRA).

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). *Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books*.