Found in Translation: Sourcing parallel corpora for low-resource language pairs

Hinrik Hafsteinsson[1,2], Steinþór Steingrímsson[1]

[1] *The Árni Magnússon Institute for Icelandic Studies*
[2] *The University of Iceland*

Abstract

This paper describes the sourcing, processing, and application of parallel text data for Icelandic and Polish for the purpose of bilingual lexicon induction (BLI), demonstrating how a parallel corpus can be compiled for a low-to-medium resource language pair that has no available parallel data, by pivoting through a common language. We show the usefulness of the corpus by training and evaluating a machine translation (MT) model on the data. Iceland's linguistic landscape is evolving, with an increasing need for multilingual support due to the growing immigrant population. Polish, in particular, stands out as the language of the largest single minority in Iceland, underscoring the importance of this project.

Keywords

parallel corpora, machine translation, data filtering, icelandic, polish

1. Introduction

Until now, openly available bilingual parallel corpora for Icelandic have been limited to English–Icelandic corpora, the most important being ParIce (Barkarson and Steingrímsson 2019). English–Icelandic data is also available in a number of multilingual data collection projects, mostly containing web-scraped texts of questionable quality. An ongoing project, 'Important Vocabulary for Multilinguality and Machine Translation'[1], aims to build resources that can be useful for second-language learners of Icelandic, training them to use the most common academic vocabulary prevalent in the teaching material used in Icelandic schools. A side-product of the project are parallel datasets for Icelandic and six other languages, English, Polish, Spanish, Tagalog, Thai, and Ukrainian, the languages of the largest immigrant communities in Iceland. Some of these resources are generated to be used for bilingual lexicon induction (BLI), expediting the compilation of small dictionaries for the academic vocabulary using the methods described for English–Icelandic BLI in Steingrímsson et al. (2022), which makes use of word alignments in parallel sentences, as well as machine translation, cross-lingual word embeddings and pivoting through available dictionaries using the approach used by Steinþór Steingrímsson, Loftsson,

DHNB 2024: Digital Humanities in the Nordic and Baltic Countries 8th Conference, May 27–31, 2024, Reykjavík, Iceland

hinrik.hafsteinsson@arnastofnun.is (H. Hafsteinsson); steinthor.steingrimsson@arnastofnun.is (S. Steingrímsson)
https://hinrik.hafsteinsson.is (H. Hafsteinsson); https://steinst.is/ (S. Steingrímsson)
0009-0003-3348-2579 (H. Hafsteinsson); 0000-0002-9776-9507 (S. Steingrímsson)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

[1] Icelandic: *Mikilvægur orðaforði fyrir fjöltyngi og vélþýðingar*

and Way (2021). For the generation of these corpora, we leverage openly available datasets to source parallel texts and lexicons.

We illustrate the methods and source datasets used for sourcing specifically Icelandic - Polish parallel text data. We select Polish as an example as Polish speakers are the largest single minority group in Iceland.

When working with parallel data for word alignment or machine translation, data quality is crucial. Poor sentence alignment and noisy data can significantly degrade both alignment quality and MT performance (see e.g. Khayrallah and Koehn (2018)). To evaluate our data quality, we test its effectiveness as training data MT systems between Icelandic and our test language, Polish."

Section 2 provides an overview of related work in the field of parallel corpora collection and filtering. Section 3 details the data collection process, including the sources of parallel texts and the filtering steps applied to the data. Section 4 outlines the training of machine translation models on the collected data, and Section 5 concludes the paper with a discussion of the results and future work.

2. Related Work

Parallel data of various quality are available for the English–Icelandic language pair. ParIce (Barkarson and Steingrímsson 2019) is specifically built for English↔Icelandic MT, with the latest version being realigned and refiltered as described in a document accompanying the corpus (Steingrímsson and Barkarson 2021). ParIce is partly a collection of parallel corpora available elsewhere, and partly data compiled specifically for the corpus, with the largest source being regulatory texts published in relation with the European Economic Area (EEA) agreement. Other prominent parallel corpora containing English–Icelandic are compiled from web-scraped data. These include Paracrawl (Bañón et al. 2020), CCMatrix (Schwenk et al. 2021), MaCoCu (Bañón et al. 2022) and HPLT (Aulamo et al. 2023). Parallel sentence pairs in the language pair are also available from multiple smaller datasets distributed on OPUS (Tiedemann and Thottingal 2020). We utilize all these datasets in our project and for training our models in the experiments described in Section 4.

To filter our data, we use a multi-step approach where both naive and more involved filters are applied to filter out poor quality sentence pairs. This roughly follows the approaches to parallel data filtering described by Steinþór Steingrímsson, Loftsson, and Way (2023) and Jasonarson et al. (2024), but they each apply a number of detailed steps which we do not.

3. Data Collection and Filtering

Considerable attention was paid to sourcing and filtering our data. We extracted our Icelandic–Polish texts from already accessible bilingual texts for English–Icelandic and English–Polish, through a process of sentence-level corpus pivoting. After extracting the Icelandic–Polish sentences, we apply our filtering steps.

3.1. Data Sources

In the case of Polish, we source our parallel texts from the following corpora:

- *CommonCrawl Aligned* (El-Kishky et al. 2020),
- *CommonCrawl Matrix* (Schwenk et al. 2021),
- *No Language Left Behind* (Costa-jussà et al. 2022),
- *OpenSubtitles* (Lison and Tiedemann 2016),
- *ParaCrawl* (versions 6 through 9, Bañón et al. 2020)
- *TildeMODEL* (Rozis and Skadiņš 2017)
- *WikiMatrix* (Schwenk et al. 2021)

All of these datasets were sourced via OPUS (Tiedemann and Thottingal 2020). They contain pre-aligned parallel texts for multiple languages, including English and Icelandic. We employ English as a pivot language in this context, which means that we first identify sentences in English that are translations of both Icelandic and Polish sentences. This approach is based on the assumption that if an Icelandic sentence and a Polish sentence have the same English translation, they are likely to convey the same meaning.

3.2. Sentence pivoting

The extraction process involves automatically comparing and matching English sentences across the Icelandic and Polish datasets. This pivoting step is not performed on each dataset in isolation. Instead, we iterate through target language datasets (here English–Polish), and compare the English sentences to a concatenation of all English–Icelandic sentence pairs found in all the datasets. This comprehensive approach aims to capture any potential Icelandic–Polish sentence pair combinations, acknowledging the method's inherent greediness. Subsequent filtering steps address any resulting redundancy or extraneous data.

To ensure optimal coverage for the available data for each language, we combine our data with "pre-pivoted" datasets, which, in the case of Polish, are MultiCCAligned (El-Kishky et al. 2020) and MultiParacrawl (Bañón et al. 2020), both of which are accessible in the OPUS catalog. This step is possibly redundant, as, in the previous step, we filtered our datasets as well, but it serves as a precaution to ensure that we do not miss any potential sentence pairs.

3.3. Data Filtering

Drawing from the comprehensive methods described in Steinþór Steingrímsson, Loftsson, and Way (2023) and Jasonarson et al. (2024), we apply a simplified approach to filtering our data, using shallow filters based on simple rules as well as score-based approaches. Initial manual checks of the output sentence pairs were performed to ensure data quality, although a comprehensive manual review of each pair was beyond the scope of this project (automatic evaluation methods are detailed in Section 4). Given the potential overlap in the input datasets, we applied a deduplication filter to the output, ensuring each sentence pair is unique in the final dataset. Additionally, we implemented a blanket-filtering step: only sentences shorter than 2000 characters were retained. As one of the main use of the sentence pairs for our project is to use word alignment for BLI, very long sentences are not suitable. Each sentence had to have at least 60% of its characters belonging to its respective language's alphabet. This criterion serves as a basic assurance that each sentence pair accurately represents its designated languages, Icelandic and Polish, respectively. Ultimately, this methodology yielded a bilingual dataset comprising 3,077,620 sentence pairs.

To gauge the quality of the sentence pairs themselves, we use Language-agnostic BERT Sentence Embedding (LaBSE, Feng et al. 2022) to vectorize the sentences and give a rough estimate on each pair's similarity. This process scores each pair from 0 to 1 based on similarity, facilitating the creation of subsets for training MT models by removing sentence pairs likely to be incorrectly translated or detrimental for MT training in some other way. The distribution of LaBSE scores for the Icelandic–Polish dataset is shown in Figure 1.

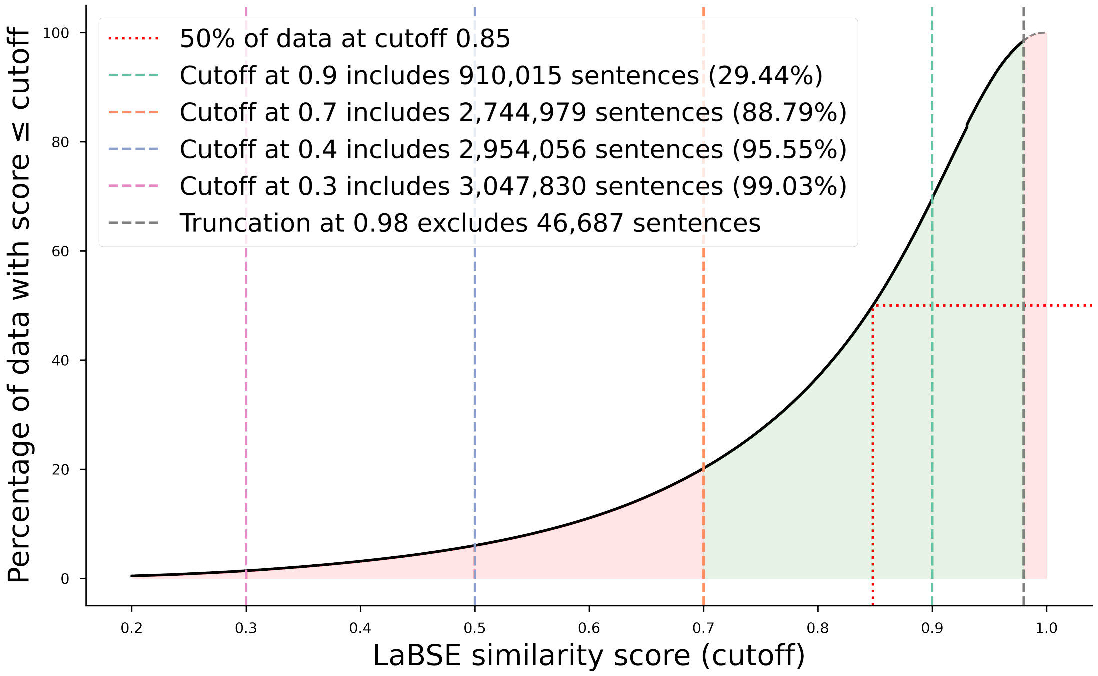

Figure 1: Cumulative Distribution of Scores for Polish–Icelandic data. Scores below 0.2 omitted from graph. Scores below 0.7 and above 0.98 are removed from dataset but included in the graph. Total sentence pairs (LaBSE 0.0-0.98): 3,077,620. The red shaded areas indicate data that is filtered out. The green shaded area indicates the data that is retained.

When gauging the optimal LaBSE cutoff value, i.e. the score above which sentence pairs are retained, the factor that must be kept in mind is that output dataset size is inversely proportional to the cutoff value. Simply put: The higher we set the cutoff value, the fewer sentence pairs we end up with. The graph shown in Figure 1 illustrates this relationship, with dashed lines indicating dataset size at cutoff levels 0.3, 0.5, 0,7 and 0.9. The optimal cutoff value is determined by the desired balance between dataset size and quality. In our case, we chose to use a cutoff value of 0.7, where any sentence pair with a lower score were removed from the dataset. This is also roughly in tune with the original LaBSE paper, where the authors suggest a cutoff value

Table 1
Comparison of BLEU scores for extant Icelandic-Polish MT models.

Model	Architecture	Parallel sentences	BLEU
Símonarson et al. (2022)	mBART-50	<unknown>	13.3
Our approach	mBART-50	2,744,979	13.0

of 0.6, based on manual inspection of English–German and English–Chinese data. Steinþór Steingrímsson, Loftsson, and Way (2023) find that for English–Icelandic a threshold of 0.7 is more useful if LaBSE is the main filtering approach.

We observe that sentence pairs in the upper extreme of the LaBSE score distribution, i.e. scores very close to 1.0, tend to be cross-language duplicates or near duplicates. Such pairs do not serve the purpose of the dataset and we thus set a hard cap at LaBSE = 0.98, removing any sentence pairs with a higher score than that. This filtering process resulted in a dataset of 2,744,979 sentence pairs, which is the final size of our new bilingual dataset.

4. MT Systems as data quality gauges

In order to try to get a measure of the quality of this new bilingual dataset, we train an MT model and evaluate it using the Flores evalution set (Goyal et al. 2021). In doing this, we assume the quality of the parallel texts gathered should be reflected in the quality of the MT system. We found that by fine-tuning the mBART-50 (Tang et al. 2020) model on the dataset using only sentence pairs that have higher LaBSE scores than 0.7, we obtain a BLEU-score (Papineni et al. 2002) of 13.0. For comparison, we evaluate the only other openly available Icelandic–Polish MT model we are aware of (Símonarson et al. 2022) on the same evaluation set and obtain a BLEU-score of 13.3, which is not a statistically significant difference as calculated using the pairwise bootstrap test (Koehn 2004). A comparison of the two models is shown in Table 1. It is promising that our baseline model is competitive with the only previously available model, an indicator of at least a good part of our sentence pairs are useful translations.

5. Conclusion and Further Work

We have described our approach to the compilation of bilingual datasets for Icelandic and languages spoken by the largest immigrant communities in Iceland. We have furthermore experimented with training an MT system for Icelandic–Polish, using MT to gauge whether our generated dataset seems to contain useful translations. Various aspects of our approach deserves further exploration. For example, while we found that most sentences having LaBSE scores very close to 1.0 were duplicates of near-duplicates, we did not inspect all these sentence pairs. And as we removed all sentences that had LaBSE > 0.98 it may be possible that we are losing some very good sentences. Other approaches to removing near-duplicates may thus be more appropriate.

Future work includes further testing on the other five languages from this project and refining our methodologies based on these initial findings. For example, additional filtering steps, or modifications to our current ones may, provide better results for the data we already have. Additionally, exploring advanced data processing techniques to handle linguistic nuances and idiomatic expressions more effectively remains a priority.

References

Aulamo, Mikko, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer van der Linde, and Jaume Zaragoza. 2023. "HPLT: High Performance Language Technologies." In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation,* 517–518. Tampere, Finland. https://aclanthology.org/2023.eamt-1.61.

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, et al. 2020. "ParaCrawl: Web-Scale Acquisition of Parallel Corpora." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 4555–4567. Online. https://aclanthology.org/2020.acl-main.417.

Bañón, Marta, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, et al. 2022. "MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages." In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation,* 303–304. Ghent, Belgium. https://aclanthology.org/2022.eamt-1.41.

Barkarson, Starkaður, and Steinþór Steingrímsson. 2019. "Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus." In *Proceedings of the 22nd Nordic Conference on Computational Linguistics,* 140–145. Turku, Finland. https://aclanthology.org/W19-6115.

Costa-jussà, Marta R, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. "No language left behind: Scaling human-centered machine translation." *arXiv preprint arXiv:2207.04672.*

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. "Language-agnostic BERT Sentence Embedding." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 878–891. Dublin, Ireland. https://aclanthology.org/2022.acl-long.62.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. "The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation."

Jasonarson, Atli, Hinrik Hafsteinsson, Bjarki Ármannsson, and Steinthór Steingrímsson. 2024. "Cogs in a Machine, Doing What They‘re Meant to Do – the AMI Submission to the WMT24 General Translation Task." In *Proceedings of the Ninth Conference on Machine Translation,* 253–262. Miami, Florida, USA. https://aclanthology.org/2024.wmt-1.18/.

Khayrallah, Huda, and Philipp Koehn. 2018. “On the Impact of Various Types of Noise on Neural Machine Translation.” In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation,* 74–83. Melbourne, Australia. https://aclanthology.org/W18-2709.

El-Kishky, Ahmed, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. “CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 5960–5969. Online. https://aclanthology.org/2020.emnlp-main.480/.

Koehn, Philipp. 2004. “Statistical Significance Tests for Machine Translation Evaluation.” In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* 388–395. Barcelona, Spain.

Lison, Pierre, and Jörg Tiedemann. 2016. “OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles.” In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC‘16),* 923–929. Portorož, Slovenia. https://aclanthology.org/L16-1147/.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. “Bleu: a Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* 311–318. Philadelphia, Pennsylvania, USA. https://aclanthology.org/P02-1040/.

Rozis, Roberts, and Raivis Skadiņš. 2017. “Tilde MODEL - Multilingual Open Data for EU Languages.” In *Proceedings of the 21st Nordic Conference on Computational Linguistics,* 263–265. Gothenburg, Sweden. https://aclanthology.org/W17-0235/.

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. “CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web.” In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 6490–6500. Online. https://aclanthology.org/2021.acl-long.507.

Símonarson, Haukur Barri, Haukur Páll Jónsson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Vilhjálmur Þorsteinsson, and Vésteinn Snæbjarnarson. 2022. “Semi-supervised Icelandic-Polish Translation System (22.09).” CLARIN-IS, http://hdl.handle.net/20.500.12537/259.

Steingrímsson, Steinþór, and Starkaður Barkarson. 2021. *ParIce: English-Icelandic parallel corpus (21.10).* CLARIN-IS. http://hdl.handle.net/20.500.12537/145.

Steingrímsson, Steinþór, Hrafn Loftsson, and Andy Way. 2021. “Pivotalign: Leveraging High-Precision Word Alignments for Bilingual Dictionary Inference.” In *Proceedings of the Workshops and Tutorials held at LDK 2021,* 190–199. Zaragoza, Spain.

Steingrímsson, Steinþór, Hrafn Loftsson, and Andy Way. 2023. “Filtering Matters: Experiments in Filtering Training Sets for Machine Translation.” In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa),* 588–600. Tórshavn, Faroe Islands. https://aclanthology.org/2023.nodalida-1.58.

Steingrímsson, Steinþór, Luke O'Brien, Finnur Ingimundarson, Hrafn Loftsson, and Andy Way. 2022. "Compiling a Highly Accurate Bilingual Lexicon by Combining Different Approaches." In *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference,* 32–41. Marseille, France. https://aclanthology.org/2022.gwll-1.6.

Tang, Yuqing, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning*. arXiv: 2008.00401. https://arxiv.org/abs/2008.00401.

Tiedemann, Jörg, and Santhosh Thottingal. 2020. "OPUS-MT – Building open translation services for the World." In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation,* 479–480. Lisboa, Portugal. https://aclanthology.org/2020.eamt-1.61.