Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3399–3405
Marseille, 11–16 May 2020

Facilitating Corpus Usage: Making Icelandic Corpora More Accessible for Researchers and Language Users

Steinþór Steingrímsson, Starkaður Barkarson, Gunnar Thor Örnólfsson
The Árni Magnússon Institute for Icelandic Studies
steinthor.steingrimsson@arnastofnun.is, starkadur.barkarson@arnastofnun.is, gunnar.thor.ornolfsson@arnastofnun.is

Abstract

We introduce an array of open and accessible tools to facilitate the use of the Icelandic Gigaword Corpus, in the field of Natural Language Processing as well as for students, linguists, sociologists and others benefitting from using large corpora. A KWIC engine, powered by the Swedish Korp tool is adapted to the specifics of the corpus. An n-gram viewer, highly customizable to suit different needs, allows users to study word usage throughout the period of our text collection. A frequency dictionary provides much sought after information about word frequency statistics, computed for each subcorpus as well as aggregate, disambiguating homographs based on their respective lemmas and morphosyntactic tags. Furthermore, we provide n-grams based on the corpus, and a variety of pre-trained word embeddings models, based on word2vec, GloVe, fastText and ELMo. For three of the model types, multiple word embedding models are available trained with different algorithms and using either lemmatised or unlemmatised texts.

Keywords: Icelandic, Corpora, KWIC, ngrams, frequency, word embeddings

1. Introduction

The need for large text corpora has become increasingly urgent in recent years. As data-oriented methods have come to dominate the field of Natural Language Processing (NLP), and in order to achieve better performance, larger datasets have to be compiled. This is especially important with the rise in popularity of neural networks and various word embedding techniques, such as *word2vec* (Mikolov et al., 2013a; Mikolov et al., 2013b), *GloVe* (Pennington et al., 2014), *fastText* (Bojanowski et al., 2017; Joulin et al., 2017) and contextual embeddings like *ELMo* (Peters et al., 2018) and *BERT* (Devlin et al., 2019).

Large corpora are also useful for other fields of research. Corpus-based linguistics is a fast-growing methodology in linguistics (see e.g. Gries, 2009), used prominently in syntax research, but also in other fields of linguistics like semantics, morphology and phonology. Lexicographers have for a long time used corpora in one form or another, mostly in the form of *citations*, but since the first dictionary based on a specific corpus, Collins COBUILD, was published in 1987 (Sinclair et al., 1987), many others have taken a similar path. In order to use corpora effectively in their work, modern lexicographers need powerful tools. They need to be able to discern between word senses, access actual usage examples, research origins of neologisms and study frequency data on word usage or usage of multiword expressions, to name a few examples. A good corpus tool is key to a comprehensive lexicographic analysis – a corpus without a good tool to access it is of little use (Kilgarriff and Kosem, 2012).

Michel et al. (2011) established a method of study they coined culturomics when launching Google N-gram viewer in 2010. By studying word usage over time, researchers can get insight into the spirit of the times, and such data can also be useful for journalists or others who may need to know when particular subjects were noticeable in public discourse and when they were not. Data on word frequency in different subcorpora can be useful in various fields of NLP, like topic modeling, machine translation or building chatbots. But such information is also important when building study material for the language classroom (see e.g. Gabrialatos, 2005).

Compiled corpora are commonly made available for download, either freely or for a price. Sometimes corpora are at the same time made available with an online KWIC tool. But it is unusual to make corpora available with a comprehensive set of tools to facilitate usage in different fields of research. The Icelandic Gigaword Corpus (IGC) is the largest existing text corpus for Icelandic. The collection work is ongoing and a new version is published every year. The corpus is freely available for download, but also accessible in different ways. It can be searched using a slightly modified version of Korp (Borin et al., 2012), word usage can be studied over time in an n-grams viewer, various frequency information can be accessed through an online frequency dictionary and pre-trained language models are available for download, to promote the use of such models in Icelandic NLP tools and projects. We introduce and describe these interconnected resources. In Section 2 we briefly depict the IGC, Section 3 describes how Korp is adapted to the Icelandic data, Section 4 outlines the n-gram viewer, the frequency dictionary is introduced in Section 5 and in Section 6 we list the language models and pre-trained word embeddings, built from the corpus data, and describe how they were trained.

2. The Icelandic Gigaword Corpus

The IGC is a collection of Icelandic texts. They are divided into subcorpora, based on the text source, and have been tokenized, pos-tagged and lemmatized. It was first published in 2018 (Steingrímsson et al., 2018) but as the collection is ongoing, a new version is published every year with more texts and reprocessed using state-of-the-art methods. The version published in 2019 contains almost 1.4 billion words. The majority of the texts, 65%, are from news media and 25% from public administration (parliamentary speeches, laws, and adjudications), see Table 1 for a full list of text types in the corpus. Almost 90% of the texts are

Subcorpora	Words	%
Newspaper Articles	896,871,188	64.85
Parliamentary Speeches	215,130,146	15.55
Adjudications	99,711,682	7.21
Transcribed Radio/Tv News	59,957,217	4.34
Sports News Websites	51,756,928	3.74
Regulations	26,924,359	1.95
Current Affair Blogs	11,916,998	0.86
Informational Articles	11,424,311	0.83
Published Books	5,247,476	0.38
Lifestyle	4,137,539	0.30
Total	1,383,077,844	100.00

Table 1: Retrieved texts for the IGC 2018

new, written in the 21st century, while most of the rest is from the last decades of the 20th century.

The texts have either a CC BY licence or a MIM licence, which was specially created for The Tagged Icelandic Corpus (MÍM) (Helgadóttir et al., 2012) and later adapted for IGC. For further discussions about licences see Steingrímsson et al., 2018.

A pipeline has been set up to collect new texts and automatically clean them, annotate and extract metadata. No manual post-editing is performed.

The annotation phase consists of sentence segmentation, tokenization, morphosyntactic tagging and lemmatization. After morphosyntactic tagging and lemmatization, the texts, together with the relevant metadata, are converted into TEI-conformant XML format (TEI Consortium, 2017). Sentence segmentation and tokenization is performed with the Reynir Tokenizer[1]. For tagging, a BiLSTM tagger (Steingrímsson et al., 2019) is used. A corpus made by concatenating the IFD corpus (Pind et al., 1991) and the MIM-GOLD corpus (Loftsson et al., 2010; Helgadóttir et al., 2012; Steingrímsson et al., 2015) were used to train the tagger and it was augmented with a morphological lexicon, The Database of Icelandic Inflections (Bjarnadóttir et al., 2019). The tagset used is a revised version of the tagset used for the IFD corpus, containing more than 670 possible morphosyntactic tags of which 559 are found in the corpus.

3. Using and Modifying Korp

Korp is a concordance search engine for large text corpora developed by researchers at the Swedish language bank, Språkbanken (Borin et al., 2012). In our work, we adapted Korp to our data, facilitating search by all the various inflectional categories of Icelandic.

3.1. Engine - CWB

Korp interacts with corpus data via the Stuttgart Corpus WorkBench (CWB) toolkit, which is designed to manage annotated text corpora of up to 2 billion words (Evert and Hardie, 2011). Queries in CWB are prompted by requests to a RESTful web API.

3.2. User Interface

Korp provides three search interfaces. A simple string literal search, a card-based search whereby you can build complex queries on all the linguistic features coded in the database, and an input for CWB's query processor where users can directly type queries.

We made various small adjustments to the UI of Korp. In addition to several cosmetic changes, we introduced search by lemma to the simple search, since the simple search interface is the most frequently used and the rich inflection of Icelandic necessitates the use of canonical forms in search. We also simplified the use of word gaps in search queries, as seen in Figure 1, due to user feedback indicating that such queries were frequent enough to warrant a dedicated UI element. Additionally, various UI features for representing data not present in the IGC were disabled, e.g. the *Map* feature which displays texts' geolocation tags on a map and a visualisation based on dependency annotation.

4. N-gram viewer

One of the services accompanying the IGC is *n-stæðuskoðari*, n-gram viewer. This is an application comparable to the Google Ngram Viewer (Michel et al., 2011; Lin et al., 2012). Our n-gram viewer is based on the NB-N-gram viewer published by the Norwegian National Library[2] (Breder Birkenes et al., 2015).

The study of human behaviour and cultural trends through quantitative analysis of digitized texts, 'culturomics', has been gaining increased popularity in recent years.

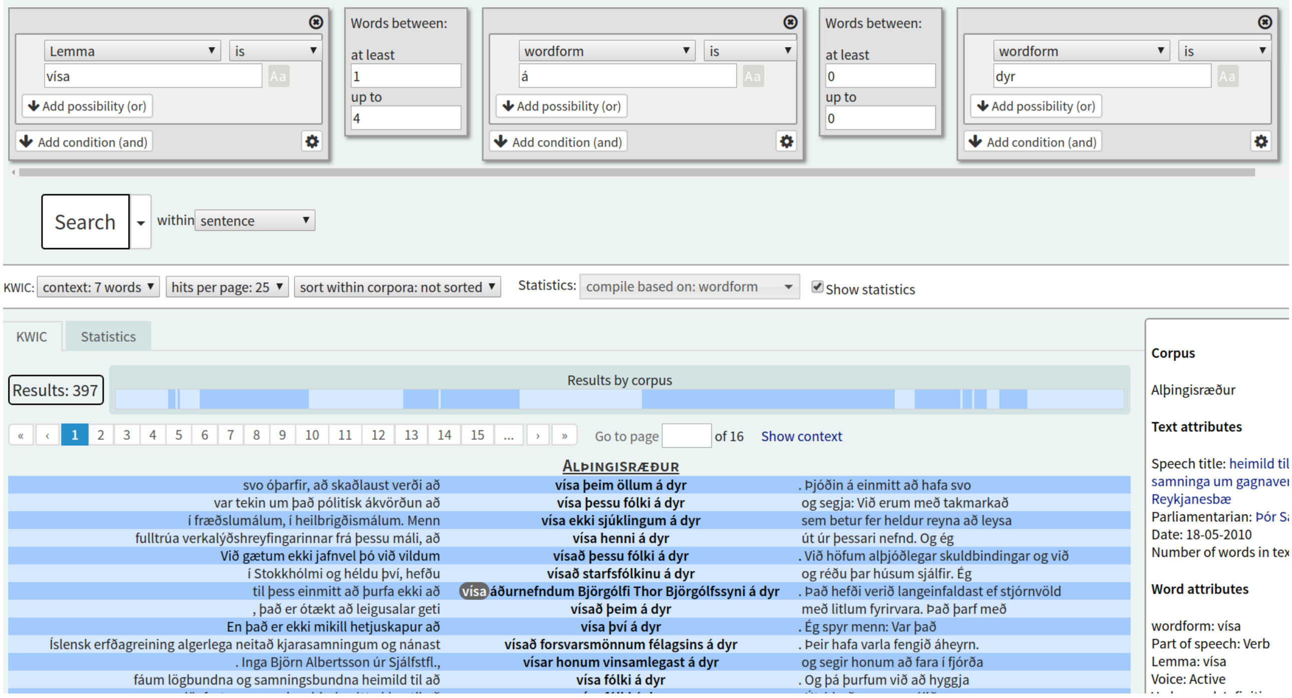
Critics of the Google Ngrams corpus have pointed out that studying large masses of text where all of them are given the same weight leads to a biased sample (Pechenick et al., 2015). They have pointed out that OCR-errors can lead to wrong results and that as all texts are treated equally, the prominence of scientific literature in recent decades leads to scientific texts being heavily sampled. We try to avoid that pitfall by allowing the users to select subcorpora from our corpus and thus focus only on certain text domains or text sources. The results are also linked to the KWIC tool for the IGC, see Section 3, so users can easily see where the counts come from and study them in context, if they so wish.

With our tool, users can chart the data by year and type of text, and see the frequency with which any word or short phrase shows up in the Icelandic Gigaword Corpus or a subcorpus thereof. The results can show how words, concepts or certain persons rise and wane in popularity over time. Although the n-gram viewer is limited in regards to the data, most of our data being from a 25-year period from the end of the 20th century and to the present year, it shows that only within this short time period there are surprising changes in word use and this can be helpful in studying current cultural phenomena, recent history and linguistics.

For the n-gram viewer, uni-, bi- and trigrams were generated from the IGC data. We created datasets for each subcorpus of the IGC, in order for the user to be able to control

[1] https://github.com/mideind/Tokenizer

[2] https://github.com/NationalLibraryOfNorway/NB-N-gram



Lemma is: vísa

Words between: at least 1 up to 4

wordform is: á

Words between: at least 0 up to 0

wordform is: dyr

Search within sentence

KWIC: context: 7 words ▾ hits per page: 25 ▾ sort within corpora: not sorted ▾ Statistics: compile based on: wordform ☑ Show statistics

Results: 397

Go to page of 16

Alþingisræður

svo óþarfir, að skaðlaust verði að	**vísa þeim öllum á dyr**	. Þjóðin á einmitt að hafa svo
var tekin um það pólitísk ákvörðun að	**vísa þessu fólki á dyr**	og segja: Við erum með takmarkað
í fræðslumálum, í heilbrigðismálum. Menn	**vísa ekki sjúklingum á dyr**	sem betur fer heldur reyna að leysa
fulltrúa verkalýðshreyfingarinnar frá þessu máli, að	**vísa henni á dyr**	út úr þessari nefnd. Og ég
Við gætum ekki jafnvel þó við vildum	**vísað þessu fólki á dyr**	. Við höfum alþjóðlegar skuldbindingar og við
í Stokkhólmi og héldu því, hefðu	**vísað starfsfólkinu á dyr**	og réðu þar húsum sjálfir. Ég
til þess einmitt að þurfa ekki að	**vísa áðurnefndum Björgólfi Thor Björgólfssyni á dyr**	. Það hefði verið langeinfaldast ef stjórnvöld
, það er ótækt að leigusalar geti	**vísað þeim á dyr**	með litlum fyrirvara. Það þarf með
En það er ekki mikill hetjuskapur að	**vísa því á dyr**	. Ég spyr menn: Var það
Íslensk erfðagreining algerlega neitað kjarasamningum og nánast	**vísað forsvarsmönnum félagsins á dyr**	. Þeir hafa varla fengið áheyrn.
. Inga Björn Albertsson úr Sjálfstfl.,	**vísar honum vinsamlegast á dyr**	og segir honum að fara í fjórða
fáum lögbundna og samningsbundna heimild til að	**vísa fólki á dyr**	. Og þá þurfum við að hyggja

Corpus

Alþingisræður

Text attributes

Speech title: heimild til samninga um gagnave Reykjanesbæ
Parliamentarian: Þór S
Date: 18-05-2010
Number of words in tex

Word attributes

wordform: vísa
Part of speech: Verb
Lemma: vísa
Voice: Active

Figure 1: An example of a card-based search with word gaps. Here we look for the phrase "vísa <x> á dyr" (e. show <x> the door), where <x> matches any sequence of 1-4 tokens.

the queries better. N-grams were created for both lowercased lemmas and word forms as they appear in the text. Bigrams and trigrams are only searchable in the viewer if they appear three times or more in the corpus. Table 2 shows counts for the different types of n-grams present in IGC on one hand and the N-gram viewer on the other hand. The n-grams in the viewer, as well as a reduced set of 4- and 5-grams, are also available for download (see Section 6.1).

N	Words	
	Total Count	Reduced Count
1	4,857,022	4,857,022
2	69,847,262	17,957,728
3	184,358,368	29,332,568
4	434,294,430	41,368,873
5	490,084,803	26,742,205
	Lemmas	
1	3,656,788	3,656,788
2	43,409,109	11,622,531
3	183,463,516	33,711,545
4	343,672,435	37,590,912
5	450,231,618	29,607,900

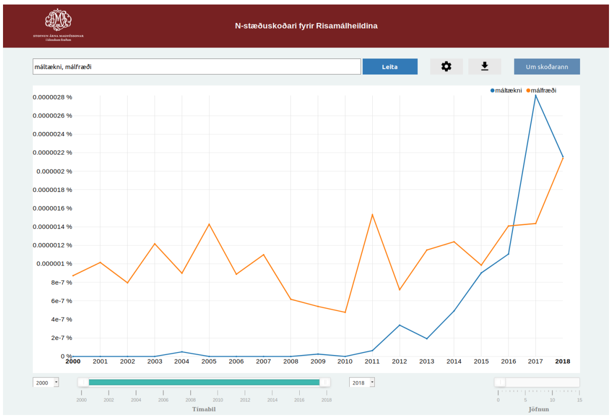
Table 2: Total count for different n-grams in the IGC. All uni-, bi- and trigrams are available for download, while the reduced set is searchable in the n-gram viewer. The reduced set of 4-, and 5-grams are available for download.

4.1. User Interface

The central element in the user interface is the chart. The chart can show the frequency of the n-grams over time given as relative frequencies (see Figure 2), or as absolute numbers. By default, the tool searches for n-grams made of word forms, but users can select an option to search for lemmas. The two are both useful in search depending on what the user is studying, as Icelandic is an inflected language. For example, when searching for unique words (unigrams), or for names, lemma search would usually give the users better results. But when searching for certain phrases, word forms would usually be a better choice.

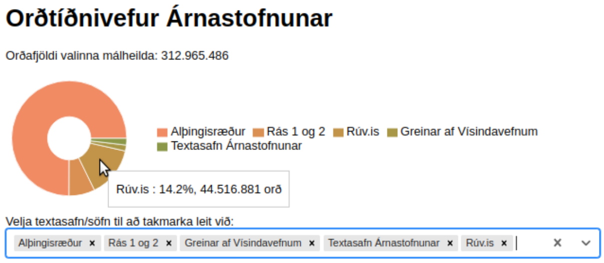
In order to give users better control over their research, they can either select the subcorpus or subcorpora to search or search all at once. This can help users deal with bias that texts from some subcorpora can have on the data.

The search functions are in line with the search functions of



máltækni, málfræði

Figure 2: Screenshot from the n-gram viewer showing relative frequencies for two unigrams, máltækni *(language technology)* and málfræði *(grammar)*.



Orðtíðnivefur Árnastofnunar

Orðafjöldi valinna málheilda: 312.965.486

Rúv.is : 14.2%, 44.516.881 orð

Velja textasafn/söfn til að takmarka leit við:

Alþingisræður × Rás 1 og 2 × Greinar af Vísindavefnum × Textasafn Árnastofnunar × Rúv.is ×

Figure 3: Visualisation of the relative token counts of selected subcorpora in the word frequency dictionary. Total token count is displayed above the pie chart, and users can examine the token counts and percentage of each subcorpus by hovering over its pie chart slice. The legend lists the selected subcorpora.

NB N-gram viewer (Breder Birkenes et al., 2015). When executing the search, users can aggregate many n-grams into one line on the chart by using the + operator. This may be useful when searching for phenomena that have more than one name or different spelling. Multiple queries, up to ten, can be executed at the same time for easier comparison. This is done by separating words or phrases with a comma. Wildcard search is also allowed. Using a wildcard in a search term will plot the ten most frequent n-grams matching the criterion. To help users interpret the results they can click on the points in the graph for each year to get a link and a query for the examples in our Korp instance, the KWIC application for the IGC (see Section 3) The n-gram viewer is accessible on `n.arnastofnun.is`.

5. The Frequency Dictionary

Word frequency is a useful metric in many fields, including linguistics, psychology and pedagogy. A word frequency dictionary for Icelandic was created in 1991, using a corpus of about 500,000 word tokens, sourced from novels, biographies and educational texts (Pind et al., 1991). Using the IGC, we have derived a new word frequency database, which spans more than 1.4 billion tokens. It contains word frequency statistics for both lexemes and inflected forms, computed for each subcorpus as well as on aggregate. Homographs are disambiguated using their respective lemmas and morphosyntactic tags.

All corpora are biased by their composition. Legal documents will not have the same word frequency distribution as a sports bulletin. This is evident in IGC, since over 90% of the text comes from either news media (64%) or public administration (26.5%). To enable users to ameliorate or investigate these biases, we allow them to limit their search queries to any set of IGC subcorpora. A pie chart of the relative sizes of the chosen subcorpora is displayed to inform the user of the composition of the corpus they are examining (See Fig. 3). The frequency dictionary is accessible on `ordtidni.arnastofnun.is`.

6. Language Models and Embeddings

In recent years neural network architectures have become state-of-the-art techniques for a range of NLP tasks, sentiment analysis (Socher et al., 2013; Kim, 2014), parsing (Dyer et al., 2015; Straka et al., 2016) and PoS-tagging (Plank et al., 2016). The first NLP model trained to work with Icelandic, a PoS-tagger, achieved better results than all previous taggers for Icelandic by a substantial margin (Steingrímsson et al., 2019). For most of these tasks, word embeddings have been shown to boost performance when used for input or additional input for neural network models. But they can be time-consuming to train and it can be difficult to compare results due to the effects of different preprocessing choices and non-determinism in the training algorithms (Fares et al., 2017). Useful pre-trained word embeddings for Icelandic, trained on large datasets, have not been available until now.

There are several different algorithms used to train word embeddings that have different pros and cons. By providing pre-trained word embeddings for Icelandic, built from the 1.4 billion word IGC we facilitate the incorporation of word embeddings in Icelandic NLP tools. We believe that will likely improve their accuracy and also make experimentation more accessible and replicable. Having readily available word embedding models also benefit students who otherwise might not have the resources to train them.

How the data is prepared for training word embeddings can affect the resulting embeddings. (Hellrich and Hahn, 2016) showed that due to the non-determinism of the embedding methods the models have reliability problems and two models trained on the same texts with the same hyperparameters can provide inconsistent results. By providing and using pre-trained embeddings the replicability problem is diminished.

We provide pre-trained models for GloVe, word2vec and FastText word embeddings. A pre-trained ELMo model will also be provided. The hyperparameters used for training are provided with the download. Information on corpus pre-processing is also provided. We also make n-grams built from the corpus available.

6.1. n-gram models

N-gram models are the simplest models that assign probabilities to sentences and sequences of words. They can be used to estimate the probability of the last word of an n-gram given the previous words or to assign probabilities to entire sentences, see e.g. (Jurafsky and Martin, 2019, Chapter 3). A variety of NLP applications take advantage of n-grams, speech recognition, handwriting recognition, spelling correction, machine translation, autocomplete features of messaging apps and more.

The n-grams created for the n-gram viewer (see Section 4) are available for download. Additionally, we have created and made available 4-grams and 5-grams that occur three times or more in the IGC. We both provide n-grams made from lemmas and word forms as they appear in the texts.

6.2. word2vec

Word2vec (Mikolov et al., 2013a) takes as its input a large corpus of text and produces vectors, with each unique word in the corpus being assigned a vector in the space. The vectors in the vector space are positioned so that words that share common traits are located close to each other. Instead of counting how a given word *w* appears close to another word *x*, a classifier is trained on a binary classification task: "Is word *w* likely to show up near word *x*?" The learned classifier weights then give us the embeddings as vectors (see e.g. Jurafsky and Martin, 2019, Chapter 6.8).

Word2vec models use one of two model architectures: continuous bag-of-words (CBOW) or continuous skip-gram. In the continuous bag-of-words architecture, the model predicts the current word from a window of surrounding words and the order of the words does not influence prediction, the context is trained on the word. In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of words, weighing nearby words more heavily than more distant ones, effectively training the word on the context (Mikolov et al., 2013a; Mikolov et al., 2013b).

We train models employing both architectures and train both on word forms and lemmas, resulting in four word2vec models.

6.3. GloVe

The main difference between word2vec and GloVe is that word2vec is a predictive model, doing incremental training by repeatedly iterating over the training corpus. GloVe, on the other hand, is global log-bilinear regression model. It creates the vectors based on ratios of co-occurrence probabilities by building a large co-occurrence matrix, which can be used to count how frequently each word is seen in a given context in the training corpus (Pennington et al., 2014). As for word2vec we train models both on word forms and lemmas.

6.4. fastText

The intuition behind fastText is that while continuous word representations, like word2vec and GloVe, trained on large unlabeled corpora can be useful for many tasks, they ignore morphology. FastText confronts that limitation by creating character n-grams from each word and learning a representation for each n-gram. One of the main advantages of the approach is that it gives better vector representations for rare words and can give representations for out-of-vocabulary words (Bojanowski et al., 2017). FastText can be trained both using a skip-gram and a CBOW model. We have pre-trained fastText models in a similar fashion as we did with word2vec, using both types of models and on both word forms and lemmas, resulting in four different embeddings models.

6.5. ELMo

The models discussed above all output just one vector for each word. This means that no matter the sense of the word in the given context, the vector is always the same. On the other hand, ELMo can generate different embeddings for a word, depending on its context. The word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on a large text corpus (Peters et al., 2018). This means that it is possible to get one vector for each word, just like with the other models, by feeding the ELMo model single words. But to take full advantage of ELMo's potentials, it should be used to generate a vector for a word in context.

Training an ELMo model using the IGC data is under way. The model will be made available for download in January 2020.

7. Availability and licensing

All datasets trained on the IGC, n-grams and word embeddings, are available under a CC BY 4.0 license. They are available for download from the IGC page on Málföng, a repository web site for Icelandic language resources[3].

The tools are open and freely accessible for everyone to use. The KWIC at `malheildir.arnastofnun.is`, the n-gram viewer at `n.arnastofnun.is` and the frequency dictionary at `ordtidni.arnastofnun.is`.

8. Future Work and Conclusion

We have described tools and datasets created to facilitate the use of the IGC corpus for a wide range of purposes and disciplines.

The IGC work is ongoing. Text collection continues, a larger version will be published every year, and as better processing tools become available the corpus is also reprocessed. Work on the tools introduced in this paper will also continue, as new methods in working with corpus data emerge and as we get feedback from our users. We have plans to add features available in Korp like *word picture*, which gives a list of the most common subjects and objects of a verb being searched, modifiers to nouns and other information on collocates. For this dependency parsing needs to be done on the data. The word embeddings will be retrained with each new version of the corpus. As we do not currently have any datasets to evaluate the accuracy of the word embeddings, creating such evaluation sets is a priority. We also have plans for building a small web application where users can explore and compare different pre-trained embeddings on-line. Recently a new type of embeddings, BERT (Devlin et al., 2019), has been outperforming all others on most downstream NLP tasks. Multilingual BERT models have been published by Google Research. They do not perform well for Icelandic, so following the lead of others, e.g. Finnish[4] and French researchers[5] (Martin et al., 2019), building BERT models for Icelandic could be a constructive step.

9. Bibliographical References

Bjarnadóttir, K., Hlynsdóttir, K. I., and Steingrímsson, S. (2019). DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, NODALIDA 2019, Turku, Finland.

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.

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp – the corpus infrastructure of språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, LREC 2012, Istanbul, Turkey.

[3] `http://malfong.is`

[4] `https://github.com/TurkuNLP/FinBERT`

[5] `https://camembert-model.fr`

Breder Birkenes, M., Johnsen, L. G., Lindstad, A. M., and Ostad, J. (2015). From digital library to n-grams: NB n-gram. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, NODALIDA 2015, Vilnius, Lithuania.

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)*, NAACL 2019, Minneapolis, Minnesota.

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., and Smith, N. A. (2015). Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL 2015, Beijing, China.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, CL 2011, Birmingham, UK.

Fares, M., Kutuzov, A., Oepen, S., and Velldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, NODALIDA 2017, Gothenburg, Sweden.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling, or wedding bells? *TESL-EJ*, 8(4):1–37.

Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass*, 3(5):1225–1241.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. (2012). The Tagged Icelandic Corpus (MÍM). In *Proceedings of SaLTMiL-AfLaT Workshop on Language technology for normalisation of less-resourced languages*, LREC 2012, Istanbul, Turkey.

Helgadóttir, S., Ágústa Svavarsdóttir, Rögnvaldsson, E., Bjarnadóttir, K., and Loftssoná, H. (2012). The tagged icelandic corpus (mim). In *Proceedings of the workshop "Language Technology for Normalization of Less-Resourced Languages" –SaLTMiL 8– AfLaT2012 at the 8th International Conference on Language Resources and Evaluation*, LREC 2016, Istanbul, Turkey.

Hellrich, J. and Hahn, U. (2016). Bad Company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, COLING 2016, Osaka, Japan.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, EACL 2017, Valencia, Spain.

Jurafsky, D. and Martin, J. H. (2019). *Speech and Language Processing (3rd Edition)*. Draft of October 16, 2019.

Kilgarriff, A. and Kosem, I. (2012). Corpus tools for lexicographers. In Sylviane Granger et al., editors, *Electronic Lexicography*, pages 31–56. Oxford University Press, Oxford, UK.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014, Doha, Qatar.

Lin, Y., Michel, J.-B., Aiden Lieberman, E., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the Google books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, ACL 2012, Jeju Island, Korea.

Loftsson, H., Yngvason, J. H., Helgadóttir, S., and Rögnvaldsson, E. (2010). Developing a PoS-tagged corpus using existing tools. In Francis M. Tyers Sarasola, Kepa et al., editors, *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, LREC 2010, Valetta, Malta.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de la Clergerie, É., Seddah, D., and Sagot, B. (2019). CamemBERT: a Tasty French Language Model. *arXiv e-prints*, page arXiv:1911.03894, Nov.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS 2013, Lake Tahoe, Nevada.

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*, 10:e0137041.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014, Doha, Qatar.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, NAACL 2018, New Orleans, Louisiana.

Pind, J., Magnússon, F., and Briem, S. (1991). Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]. *The Institute of Lexicography, University of Iceland, Reykjavik, Iceland.*

Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of the* 54^{th} *Annual Meeting of the Association for Computational Linguistics*, ACL 2016, Berlin, Germany.

Sinclair, J., COBUILD (Information retrieval system), University of Birmingham. Department of English Language and Literature, and Collins Publishers. (1987). *Collins COBUILD English Language Dictionary*. Collins, London, England.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2016, Seattle, Washington.

Steingrímsson, S., Helgadóttir, S., and Rögnvaldsson, E. (2015). Analysing Inconsistencies and Errors in PoS Tagging in two Icelandic Gold Standards. In *Proceedings of the* 20^{th} *Nordic Conference of Computational Linguistics*, NODALIDA 2015, Vilnius, Lithuania.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. (2018). Risamálheild: A very large Icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan.

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 Recent Advances in Natural Language Processing*, RANLP 2019, Varna, Burgaria.

Straka, M., Hajic, J., and Straková, J. (2016). Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016, Portorož, Slovenia.

TEI Consortium, e. (2017). Tei p5: Guidelines for electronic text encoding and interchange. 3.2.0. last updated on 10th july 2017.