Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2371–2381
Marseille, 20-25 June 2022

Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus

Starkaður Barkarson, Steinþór Steingrímsson, Hildur Hafsteinsdóttir
The Árni Magnússon Institute for Icelandic Studies
{starkadur.barkarson, steinthor.steingrimsson, hildur.hafsteinsdottir}@arnastofnun.is

Abstract

The Icelandic Gigaword Corpus was first published in 2018. Since then new versions have been published annually, containing new texts from additional sources as well as from previous sources. This paper describes the evolution of the corpus in its first four years. All versions are made available under permissive licenses and with each new version the texts are annotated with the latest and most accurate tools. We show how the corpus has grown almost 50% in size from the first version to the fourth and how it was restructured in order to better accommodate different meta-data for different subcorpora. Furthermore, other services have been set up to facilitate usage of the corpus for different use cases. These include a keyword-in-context concordance tool, an n-gram viewer, a word frequency database and pre-trained word embeddings.

Keywords: text corpora, Icelandic, gigaword

1. Introduction

Large text corpora are useful in various fields of research. They can be used to study language variation (Iomdin et al., 2013), for compiling dictionaries, see e.g. Sinclair (1987), Gizatova (2016), Jónsdóttir and Úlfarsdóttir (2020), and for developing NLP applications.

Since the advent of Word2Vec (Mikolov et al., 2013), static word embeddings trained using neural networks, various neural network language models have been employed for most if not all common NLP tasks. These language models require large corpora of target language texts for training. Although the first word embedding models could produce meaningful results with only a few million tokens in the training data (Jungmaier et al., 2020), for best results they are trained on corpora containing hundreds of millions or even billions of words (Grave et al., 2018).

While ever larger language models are being developed, with GPT-3 (Brown et al., 2020) trained on approximately 500 billion tokens and Gopher (Rae et al., 2021) using two trillion words for training, the more common BERT-like models (Devlin et al., 2019) of contextualized embeddings, traditionally trained on multibillion token corpora, have been shown to give very competitive results when trained on only approximately one billion tokens, see e.g. Martin et al. (2020).

The need for large and updated text corpora is thus evident if a language is to be a part of the development of language technology (LT). In 2019, a national language technology programme for Icelandic (Nikulásdóttir et al., 2020) was started. As a part of that programme, the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018) was to be enlarged and evolved, with new versions published every year.

This paper describes the evolution of the corpus from its initial version, published in 2018, to the fourth version, published in 2021. We describe the content, annotation process and licensing, the division of the data into independent subcorpora, structure of the published data, evaluation of POS-tagging accuracy, tools for doing research using the data, and finally we give examples of how the corpus has been used in these first four years.

The corpus can be accessed and used in various ways. All versions of the IGC are made available for download with permissive licenses. They can also be explored using the corpus research tool KORP (Borin et al., 2012) which employs the IMS Corpus Workbench (Evert and Hardie, 2011) for indexing and searching the corpora. The texts have also been processed so the word usage can be analysed in an n-gram viewer, as well as an online frequency dictionary. Word embedding models trained on Word2Vec and GloVe using the corpus have also been made available.

2. Evolution of the Corpus

The IGC project started in 2017, with the aim to compile as large a corpus as possible with the minimum amount of work and resources (Steingrímsson et al., 2018). The corpus should contain more than a billion running words from contemporary texts, morphosyntactically tagged and lemmatized, and provided with metadata. Only digitally available texts were to be included and formats that might pose a difficulty were not processed. A new version was planned to be published every year and the corpus was to be clearly versioned in order to facilitate reproducible experiments.

2.1. Content

The first version of the IGC was published in 2018, with texts dating until December 2017. As shown in Table 1 it contained approximately 1,259 million running words. Around 72% of the texts were sourced from news media, 26% were from official documents, the biggest part parliamentary speeches, but also laws and adjudications, while the rest, less than 2%, was from other sources, e.g. published books, Wikipedia and the Icelandic web of science – a website containing questions and answers for the public on all aspects of science.

The version published in the following year, 2019, grew in size by roughly 11%. Its composition did not change much except that text from six new news media were added and adjudications from a new judicial level, The Court of Appeal. The increase was mainly due to additional texts from the year before, from previously available sources, as

Version	Year of publication	Words (M)	Number of sources	POS-tagger	Tokenizer	MIM-GOLD Tagset
IGC-2017	2018	1,259	54	IceStagger	IceNLP	v. 1.0
IGC-2018	2019	1,394	60	IceStagger	IceNLP	v. 1.0
IGC-2020	2020	1,532	73	ABLTagger 0.9	Tokenizer	v. 2.0
IGC-2021	2021	1,880	108	ABLTagger 2.0.4	Tokenizer	v. 2.0

Table 1: The number of running words (millions) and sources in the four published versions of the IGC, as well as information on annotation tools and tagset. There is no version with the suffix 2019 due to a change in naming conventions. The first two versions refer to the year of the most recent texts, but since then the year of publication has been used.

well as better scraping techniques adding some new text from earlier years.

In 2020, twelve news media sources were added, mostly from smaller publications. The increase in number of words was around 10%.

With the publication of the 2021 version, which contains approximately 1,880 million running words, an increase of 22%, it was decided to diversify the corpus, add text from new domains (social media, scientific journals, and resolutions and bills from the parliament) and try to increase the ratio of smaller domains (books). In order to do that the rule of working only with digitally available texts and exclude formats that might pose difficulty had to be abandoned. With these changes it was also decided to restructure the corpus and split it into eight subcorpora (see Section 3).

2.2. Annotation

The annotation phase consists of sentence segmentation, tokenization, morphosyntactic tagging and lemmatization.

The two first versions were tokenized with the IceNLP package (Loftsson and Rögnvaldsson, 2007) but the latest two with Tokenizer[1]. A manual inspection of their output gave reason to choose the latter over the IceNLP tokenizer.

IceStagger (Loftsson and Östling, 2013), which was used to tag the first two versions of the IGC, was trained on two gold standards combined, MIM-GOLD 1.0 (Loftsson et al., 2018) and IFD 2018.10 (Helgadóttir, 2018), and augmented with data from DMII (Bjarnadóttir, 2012). Its accuracy, when tested on the gold standards, is 93,71% (Barkarson, 2017). Since the 2020 version, the corpus has been tagged with different versions of ABLTagger (Steingrímsson et al., 2019), IGC-2020 with version 0.9 (95.15% reported accuracy) and IGC-2021 with version 2.0.4 (95.78% reported accuracy). ABLTagger was trained with a new version of the two gold standards: IFD 2020.05 (Helgadóttir et al., 2020) and MIM-GOLD 20.05 (Barkarson et al., 2020a). These new versions use a slightly different tagset and have undergone manual correction from the previous version. The tagset contains about 700 possible tags of which 538 appear in IGC[2].

All the versions of the IGC were lemmatized using Nefnir (Ingólfsdóttir et al., 2019) which is reported to obtain an accuracy of 99.55% when tokenizing correctly tagged text.

In the 2021 version, anonymization of names was carried out for adjudications by using named entity recognition features of GreynirSeq[3] to find names of people, and replace the names with a string that indicates the gender and case of the name.

2.3. Permission Clearance and Licensing

One of the design considerations for the IGC was to make the corpus available with a permissive license, such as a Creative Commons license[4]. However, Creative Commons licensing is not widely known in Iceland and some text providers were hesitant to agree to such an open licence, so eventually it was necessary to use a more restricted license for some of the news texts and for published books. The licence selected was the one developed for the Tagged Icelandic Corpus (MIM) (Helgadóttir et al., 2012), released in 2013. Both licenses allow use of the data for all research and language modelling.

As of the third version of the IGC, the MIM-licence has not been used for any new data sources except for published books. All new providers of news media have shared their data under CC BY 4.0 and some providers who chose to use the MIM licence previously have agreed to allow their data to be published under CC BY 4.0.

For some corpora the original texts were divided into smaller parts and shuffled for copyright compliance.

3. Icelandic Gigaword Corpus Divided into Eight Corpora

The fourth version of the IGC is in many respects different from the previous versions. At the same time as an effort was made to reduce the bias towards news and public data, by adding new sources, the corpus was divided into eight individual corpora. The biggest challenge was to find a format for the eight corpora that would be uniform, but at the same time allow for diversity in terms of metadata and the structure of the data.

We begin by describing the subcorpora and then describe their structure and format.

3.1. Description of the Eight Constituents of the IGC

As shown in Table 2 the corpus is still heavily biased towards texts from news media (64.38%) and public texts (16.50%). The addition of texts from books and journals

[1] https://pypi.org/project/tokenizer/

[2] Description of the tagset can be found in the MIM-GOLD 20.05 package.

[3] https://github.com/mideind/GreynirSeq

[4] https://creativecommons.org/

	IGC-2017		IGC-2018		IGC-2020		IGC-2021	
Subcorpora	**Words**	**Ratio (%)**	**Words**	**Ratio (%)**	**Words**	**Ratio (%)**	**Words**	**Ratio (%)**
News	901,002,839	71.56	1,022,626,345	73.34	1,146,169,569	74.83	1,209,940,022	64.35
Social Media	11,886,951	0.94	11,998,988	0.86	12,073,300	0.79	319,959,250	17.02
Parliamentary	210,490,367	16.72	215,280,201	15.44	220,798,376	14.41	222,531,059	11.84
Adjudications	92,702,603	7.36	100,345,365	7.20	106,145,637	6.93	56,595,797	3.01
Laws	27,079,422	2.15	27,256,120	1.95	27,346,134	1.79	40,612,997	2.16
Journals	4,617,751	0.37	4,885,110	0.35	5,068,581	0.33	17,211,891	0.92
Books	5,199,934	0.41	5,252,601	0.38	5,201,312	0.34	13,340,865	0.71
Wikipedia	6,174,619	0.49	6,787,384	0.49	8,963,253	0.59	0	0

Table 2: The size and ratio of each of the four published versions of the IGC.

did not do much to change that - they are still the smallest subcorpora, both with under 1% of the total of the IGC - but newly added texts from the social media have changed the ratio a lot since texts from news media and the public domain now only add up to 81% instead of 98% before. Data from Wikipedia, which has been part of the IGC from the start, was not included in IGC-2021 but will be part of a new corpus with encyclopedic data for the next version.

While the IGC contains some older texts the IGC can be considered a corpus of contemporary texts. Almost 73% of IGC-2021 are texts from 2000 and later, 6.5% are from before 1980, and only 0.003% from before 1900. The dates for each constituent of the IGC can be found in Table 3, along with other key information.

3.1.1. IGC-News

Two of the constituent corpora contain text from news media, dating from February 1998 to the end of year 2020. IGC-News1 (Barkarson and Steingrímsson, 2021a) contains texts with CC-BY licence and IGC-News2 (Barkarson and Steingrímsson, 2021b) contains texts with restricted licence (MIM). Each corpus contains several subcorpora since texts from each media is saved as a unique corpus. Some of the news sites we collect data from publish news in English and Polish as well as Icelandic. They are a very small part of the published data and were not filtered out before the 2021 version was published. From that version on we automatically check the language of all texts and filter out texts that are not in Icelandic.

3.1.2. IGC-Social

IGC-Social (Barkarson et al., 2021a) contains three subcorpora.

Forums: IGC-Social-Forums contains text from two online forums, each stored as an individual subcorpus. The texts from each month were grouped together and the sentences within each month reshuffled. While this presents a disadvantage for certain types of research, e.g. discourse analysis, this removes copyrights issues and all information about authors.

Blogs: IGC-Social-Blogs is divided in three parts, one for each blog site included. These were already incorporated in the previous versions of the IGC.

Tweets: IGC-Social-Tweets contains tweets in Icelandic. The tweets were grouped by month and no information about users nor exact dates are included. Due to Twitter's licensing resrictions, the files have been "dehydrated", in other words all texts have been removed but information about tweet IDs and POS-tags are still present so users can rehydrate the corpus (insert the text again) by using Twitters API. We distribute the corpus with necessary data and tools to do so in a simple way.

3.1.3. IGC-Parla

IGC-Parla (Barkarson and Steingrímsson, 2021c) contains all parliamentary speeches available on the Althingi website in edited format[5]. The oldest speech is from 1911 but the data is quite sparse until the year 1923. The vast majority of speeches given from the mid-20th century until the end of 2020 are included in the corpus.

3.1.4. IGC-Laws

IGC-Laws (Steingrímsson and Barkarson, 2021) contains three subcorpora. All the data is sourced from the Althingi website.

Proposals and resolutions: IGC-Laws-Proposals contains proposals and resolutions submitted to Alþingi, dating from November 1988.

Bills: IGC-Laws-Bills contains explanatory reports and observations that are attached to bills that have been submitted to Althingi since October 1988.

Laws: IGC-Laws-Laws contains current Icelandic law as of the date the corpus is published.

3.1.5. IGC-Adjud

IGC-Adjud (Steingrímsson and Barkarson, 2021) contains three subcorpora, one for each judicial level: the district courts, the Court of appeal and the Supreme Court. As shown in Table 2 the content of IGC-Adjud shrunk in size since former versions as references in the documents from the supreme court to documents from the district courts were omitted to avoid duplication.

3.1.6. IGC-Books

The first version of the IGC contained texts from 114 books by authors who had given permission for the usage of their books for a previous corpus project (Helgadóttir et al., 2012). For the 2021 version, a lot of effort was put into obtaining permissions from publishers and authors, to gather

[5] www.althingi.is

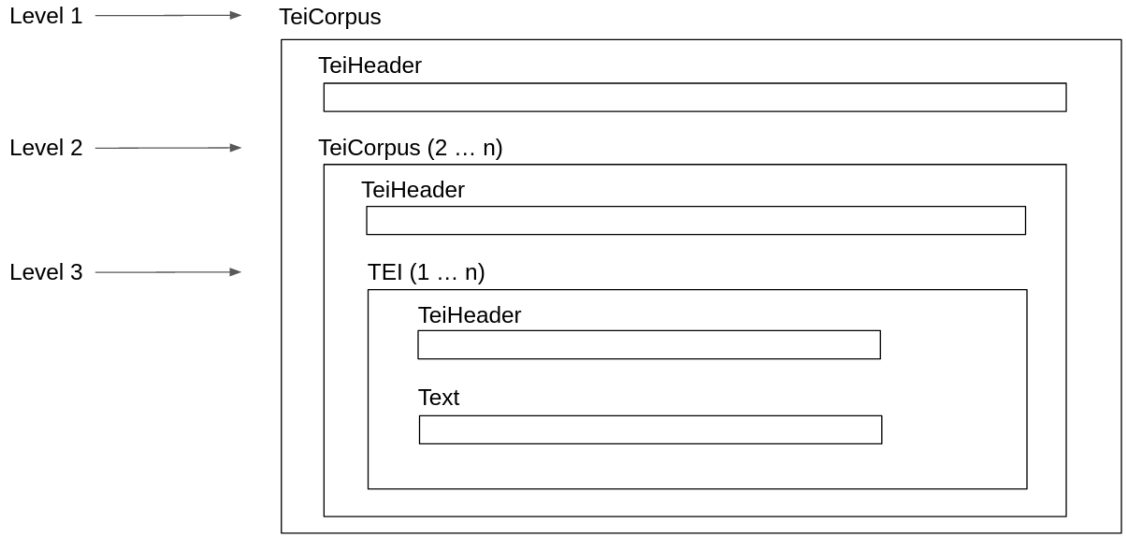

Figure 1: The structure of a typical constituent of the IGC. The root is a teiCorpus element (level 1) which contains a teiHeader element with general information about the corpus. This corpus has two or more subcorpora, (level 2) which are located in another teiCorpus elements. Each subcorpus contains a teiHeader element with metadata, specifically for that subcorpus, and several TEI elements containing the metadata and the text of one specific text (article, book, etc.). Each TeiCorpus and TEI element is located in an individual file which is referred to from the parent element, to avoid having too large files.

more texts from published books and extract the texts from PDF and desktop publishing documents. This proved to be a slow and not so rewarding process. IGC-Books (Barkarson et al., 2021c) contains texts from 351 books, published from 1968 to 2020. In order to adhere to publisher's requests, texts were split into parts that did not include more than 500 words and reordered in a random way. Paragraphs were not dismantled unless they contain more than 500 words. If a text contained less than 5000 words, sentences of each paragraph were also rearranged.

3.1.7. IGC-Journals

IGC-Journals (Barkarson et al., 2021b) contains articles from scientific and scholarly journals, as well as texts from websites containing scientific or scholarly articles, including those from the Icelandic web of science that were already present in the older versions of the ICC. As with the books, it was time consuming to obtain both permissions and the texts. The corpus contains texts from 14 printed journals and four websites. The sentences of each article were reshuffled.

3.2. Structure of the TEI-files

The first three versions of the IGC were published in two parts, depending on the licence of the texts. They were distributed in a TEI-conformant XML format [6], but full advantage was not taken of the richness of the TEI-format. Although the corpus contained texts from different domains, and each domain was placed in a separate folder, the structure of the xml-files did not suggest any subcorpora. This resulted in poor meta-data for some categories and a lack of overall information about the texts. For IGC-2021 the structure of the TEI-files was changed thoroughly. Although the corpora differ both in terms of metadata and licence restrictions, all the corpora have the same overall structure, which is illustrated in Figure 1, and the XML files have a similar structure of both the teiHeader element and the text element. The text element of IGC-Parla deviate the most from the common structure but there we followed the Parla-CLARIN schema[7].

Each corpus is distributed in two variants: the first is the fully marked-up corpora, but with plain text, for those who want to have the text untokenized, while the second is identical to the first one, but with added linguistic annotations (POS-tags and lemmas) to the texts. Each corpus is nested in a teiCorpus element with a teiHeader that contains the metadata. If the corpus has no subcorpora (IGC-Parla, IGC-Books, IGC-Journals) then the teiHeader is followed by a TEI element for each text (article, book, bill, etc.). The TEI element also contains a teiHeader element with information about the text and text itself in the text element.

In most cases the corpora contain subcorpora. As an example, IGC-News has each media as one subcorpus and IGC-Laws is divided into bills, laws and proposals. In that case the structure is identical to the one illustrated in Figure 1 where the teiHeader element of the root is followed by several teiCorpus elements that contain the subcorpora.

It is not feasible to store a whole corpus in one file, since it would be far too big. For that reason, and to make it easier to access the subcorpora, each TeiCorpus and TEI element is located in an individual file which is referred to from the parent element. Since components of the TEI-element will always refer to the metadata contained by the closest parent, each xml-file, containing a teiCorpus element, can stand as an independent root-file.

[6] https://tei-c.org/Guidelines/P5/

[7] https://github.com/clarin-eric/parla-clarin

Constituent Corpora of IGC 2021						
Corpus	**Subcorpus**	**Sources**	**No. texts**	**Words**	**Dates**	**Licence**
IGC-News						
	IGC-News1	24	16,561,21	354,459,688	2001-2020	CC BY 4.0
	IGC-News2	24	3,077,992	855,480,334	1998-2020	MIM
IGC-Social						
	IGC-Social-Forums	2	1,312,143	181,404,769	2000-2020	CC BY 4.0
	IGC-Social-Blogs	3	38,349	8,967,128	1973-2018	CC BY 4.0
	IGC-Social-Tweets	1	10,168,408	129,587,353	2007-2020	CC BY 4.0
IGC-Parla		1	403,746	212,873,555	1911-2020	CC BY
IGC-Adjud						
	IGC-Adjud-[Regional]	1	12,617	43,752,014	2020-2019	CC BY 4.0
	IGC-Adjud-[Appeals]	1	1,823	2,087,667	2018-2020	CC BY 4.0
	IGC-Adjud-[Supreme]	1	11,802	10,756,116	1999-2020	CC BY 4.0
IGC-Laws						
	IGC-Laws-Proposals	1	6,585	12020955	1998-2020	CC BY 4.0
	IGC-Laws-Bills	1	6,975	25,974,010	1988-2020	CC BY 4.0
	IGC-Laws-Laws	1	837	2,618,032	1275-2020	CC BY 4.0
IGC-Journals		20	19,805	17,211,891	1979-2020	CC BY 4.0
IGC-Books		27	351	13,340,865	1968-2020	MIM

Table 3: Key information for the constituent parts of IGC 2021, and their subcorpora.

4. Evaluation

When comparing taggers, the accuracy of a tenfold cross-validation over a gold standard is a good indicator. However, it does not necessarily tell us much about how well they do when tagging texts from a certain domain. The ratio of literary texts is for example much higher in IFD and MIM-GOLD, the gold standards that were used to train the taggers, than in the IGC. Conversely, the ratio of news media texts is much higher in the IGC than in the gold standards.

In order to predict the accuracy of the POS-tagging for different domains we created a bundle of evaluation sets, one set for each domain (Barkarson et al., 2020b). The evaluation set contains texts extracted from the 2020 version of IGC, when tweets, texts from discussion forums and scientific journals had not yet been included in the corpus. Therefore, evaluation sets for these texts have not been compiled. Texts were selected randomly from nine different domains, and four complementary methods used to flag tags that should be checked manually.

1. We started by tagging the test set with three different taggers: ABLTagger, IceStagger and TriTagger, a re-implementation of the statistical tagger TnT (Brants, 2000) and a part of the IceNLP package (Loftsson and Rögnvaldsson, 2007). A script flagged all the cases where the three taggers did not agree.

2. We used the Decca software package[8] to find identical strings of words (5-15 words) that were tagged differently.

3. We used IceParser (Loftsson and Rögnvaldsson, 2007) to perform shallow parsing on the corpus and find disagreement errors within noun phrases.

4. Finally, we looked at all tokens tagged with **e** (foreign words), **x** (unanalyzed words) and **n—-s** (proper nouns with no further analyses), which we found to be commonly mistagged, and where the tag is **kt** (short form of nouns) while the word starts with a capital letter, indicating it should possibly be tagged as a proper noun.

In total, 19,474 tags (19.23% of the set) were manually checked. 4,652 tokens, or 4.59% of the whole set, were corrected, or almost 24% of the tokens that were flagged and checked. When incorrect tagging was due to wrong tokenization, we corrected both the tokenization and the tag. In most cases, those were abbreviations at the end of a sentence where the dot had been split from the abbreviation.

Finally, we used the sets to evaluate the accuracy of the two taggers that had been used to tag the IGC, ABLTagger 2.0.4 and IceStagger. The results are shown in Table 4.

The tagging accuracy using ABLTagger is considerably higher than when IceStagger is used. In some cases, accuracy goes up by more than 3%. In these cases, much of the gain comes from ABLTagger's ability to resolve case and gender better than the older tagger, and because ABLTagger is better at handling long-distance dependencies. Parliamentary speeches and legal texts have somewhat lower gains in accuracy. The largest reason for that is the prevalence of abbreviations unique to these texts. This also seemst to affect the tagger's ability in tagging the surrounding words. The only domain where we don't see much gain in accuracy is in the Books corpus. Upon inspection many of the incorrectly tagged words are out of vocabulary words such as foreign names and uncommon forms of place names rarely used, which are far more common in this evaluation set than in the others.

[8] http://decca.osu.edu/

Domain	Tokens	ABL	IceStagger
Adjudications	12,442	96.79	92.61
Books	10,353	94.86	94.83
Educational websites	10,772	95.46	93.88
Legal texts	12,177	95.91	94.51
Blogs	11,662	96.78	93.59
Parliamentary speeches	12,358	94.34	93.64
News (web and print)	11,459	96.81	93.55
Sports news	9,272	95.47	92.67
TV and radio news	10,766	97.61	94.58
TOTAL	101,261	96.02	93.60

Table 4: The accuracy of ABLTagger 2.0.4 and IceStagger when used to POS-tag nine different domains from IGC.

5. Examples of Usage

Since the publication of the first version of IGC in 2018, the corpus has shown itself to be a valuable resource, both for building LT tools of Icelandic as well as for linguistics research. Below a few examples of different projects that have used the corpus are listed.

- 50,000 most frequent lemmas in the IGC were used to add to the vocabulary of DMII Core (Bjarnadóttir et al., 2019)
- Texts from the the IGC were used during the crowdsourcing data collection for Icelandic speech recognition (Mollberg et al., 2020)
- ALEXIA (Friðriksdóttir et al., 2021), a lexicon acquisition tool designed to facilitate the compilation and expansion of lexical databases and dictionaries, used the IGC as a default corpus.
- The text from the parliamentary speeches were used when compiling ParlaMint-IS, a subcorpus of the ParlaMint project (Erjavec et al., 2021).
- Texts from the news domain in IGC were used to create a synthetic parallel corpus, by way of backtranslations, for training machine translation models (Símonarson et al., 2020).
- Texts from the corpus were parsed to create The Icelandic Contemporary Treebank (Arnardóttir et al., 2020).
- Texts from the corpus were used to train word embeddings for evaluating Icelandic versions of Multi-SimLex and IceBATS, described in Friðriksdóttir et al. (2022).
- Study of the effect of case syncretism of the acceptability of a variety of syntactic constructions. The corpus is used to find examples of constructions (Snorrason, 2021; Sigurðsson and Wood, 2021).
- A large part of the research project "Ditransitives in Insular Scandinavian", directed by Jóhannes Gísli Jónsson and Cherlon Ussery and funded by the Icelandic Research Fund, is dedicated to collecting data in the IGC (Magnússon, 2019; Jónsson, 2020).

6. Related Work

Before the publication of the IGC, the largest corpus existing for the Icelandic language was The Tagged Icelandic Corpus (MIM) (Helgadóttir et al., 2012). It was released in 2013 and contained 25 million running words from various genres dating from the first decade of the 21st century.

The first Gigaword corpus was the English Gigaword (David Graff, Christopher Cieri, 2003). It was produced by the Linguistic Data Consortium (LDC) and first published in 2003, and consisted of roughly one billion words of English-language newswire texts. Since then similar corpora, i.e. collections of comprehensive amount of unannotated newswire texts, have been published by LDC in other languages, such as Chinese, Arabic, French and Spanish. A more recent initiative is the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021) that unlike the aforementioned corpora covers a wide array of domains, with news texts being only about 40 million of over one billion words.

7. Availability and Use

All the previous versions of the IGC as well as the eight constituent parts of IGC-2021, are available for download from the CLARIN-IS repository[9] For other uses, such as linguistics and lexicography research, teaching or other studies, the data is available in a concordance tool run on KORP[10], where the latest version as well as all previous versions are accessible for search and research. Users of the search interface can take advantage of the annotation of the texts when specifying search criteria.

The corpus texts are also available through an n-gram viewer and a word frequency database. The word frequency database[11] contains word frequency statistics for both lexemes and inflected forms, computed for each subcorpus as well as on aggregate, with homographs being disambiguated using their respective lemmas and morphosyntactic tags. The n-gram viewer[12], is based on NB N-gram viewer (Birkenes et al. 2015). It allows the user to chart the data by year and type of text, and shows the frequency with which any word or short phrase shows up in the IGC or any of its subcorpora. These additional services are described in more detail in Steingrímsson et al. (2020).

Information on the corpus and its availability is kept up-to-date on a dedicated website[13]. To aid researchers, students or developers of LT tools using the corpus, the information site also has downloads for n-grams (n up to 5). Word embedding models trained and evaluated on IceBATS (Friðriksdóttir et al., 2021) as described in Friðriksdóttir et al. (2022) are also available with accompanying metadata, FastText models (Friðriksdóttir et al., 2022a), GloVe models (Friðriksdóttir et al., 2022b) and Word2Vec models (Friðriksdóttir et al., 2022c).

[9] http://repository.clarin.is
[10] https://malheildir.arnastofnun.is
[11] https://ordtidni.arnastofnun.is
[12] https://n.arnastofnun.is
[13] http://igc.arnastofnun.is

8. Conclusion and Future Work

In the four years since the first publication of the IGC the corpus has grown from 1,259 to 1,880 million running words, almost 50%. For the next publication in 2022, another discussion forum will be added to IGC-Social, containing approximately 380 million words. Work on gathering more books for the book corpus and more journals for the journals corpus will continue for that version, as well as a new corpus with encyclopedic data will be added, containing data from the Icelandic Wikipedia and possibly other related sources. Adding the annual estimated growth of other domains, the size will likely pass 2,300 million running words in the fifth version of the corpus.

Other future plans include annotating the corpus with Universal Dependencies (UD) tags as well as named entities. Data for training automatic tools for such annotation have been made available for Icelandic, a UD treebank containing 158K words from contemporary texts are available in the Universal Dependencies project (Zeman et al., 2021) and the MIM-GOLD corpus has been manually annotated with named entities (Ingólfsdóttir et al., 2020).

As the division of the corpora into eight subcorpora, which may have different metadata and sometimes different structure of the textual data, requires more effort from the user when gathering and processing the texts, and due to the fact that the Twitter data has to be rehydrated, we aim to publish a python package that can simplify the work for the user by allowing for simple download and processing of the different subcorpora.

In the first four years, a new version of the IGC was published each year. With the end of the Icelandic language technology programme, which has funded a large part of the work on the corpus, it is probable that the growth of the IGC will slow down. The division of the IGC into eight subcorpora makes it possible to publish only one or a few of these at a time. Subcorpora in domains that grow substantially in size every year (news, parliamentary speeches, social media) are planned to be published annually, while others, like books and journals, may be published with longer intervals. If tools used for annotation do not change between years, it may also be an option to publish only additional data in some years, instead of republishing the whole corpus. As well as the latest version, all previous versions of the corpus will be available for users to download and within KORP on malheildir.arnastofnun.is.

9. Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019–2023. The programme, which is managed and coordinated by Almannarómur[14], is funded by the Icelandic Ministry of Education, Science and Culture.

10. Bibliographical References

Barkarson, S. (2017). Þjálfun málfræðimarkarans stagger með nýjum gullstaðli. Master's thesis, University of Iceland.

Bjarnadóttir, K. (2012). The Database of Modern Icelandic Inflection. In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages -SaLTMiL 8 – AfLaT2012*, pages 67–72. Istanbul, Turkey.

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*, pages 146–154, Turku, Finland.

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'12)*, pages 474–478, Istanbul, Turkey.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231, Seattle, Washington, USA.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

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.

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

Friðriksdóttir, S. R., Jasonarson, A., Steingrímsson, S., and Sigurðsson, E. F. (2021). ALEXIA: A Lexicon Acquisition Tool. In *Proceedings of CLARIN Annual Conference 2021*, pages 64–67, Online.

Friðriksdóttir, S. R., Daníelsson, H., Steingrímsson, S., and Sigurðsson, E. F. (2022). IceBATS: An Icelandic Adaptation of the Bigger Analogy Test Set. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Gizatova, G. (2016). A Corpus-based Approach to Lexicography: Towards a Thesaurus of English Idioms. In *Proceedings of the 17th EURALEX International Congress*, pages 348–354, Tbilisi, Georgia.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Helgadóttir, S., Ásta Svavarsdóttir, Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. (2012). The Tagged

[14] https://almannaromur.is/

Icelandic Corpus (MÍM). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages -SaLTMiL 8 – AfLaT2012*, pages 67–72, Istanbul, Turkey.

Ingólfsdóttir, S. L., Loftsson, H., Daðason, J. F., and Bjarnadóttir, K. (2019). Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 310–315, Turku, Finland.

Iomdin, B., Piperski, A., and Somin, A. (2013). Linguistic Problems Based on Text Corpora. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 9–17, Sofia, Bulgaria.

Jónsdóttir, H. and Úlfarsdóttir, Þ. (2020). Omdannelsen af en flersproget til en monolingval ordbog. *Nordiske Studier i Leksikografi*, (15).

Jónsson, J. G. (2020). Object inversion in Icelandic and the Risamálheild Corpus. In Kristin Hagen, et al., editors, *Bauta: Janne Bondi Johannessen in memoriam.* Oslo Studies in Language 11(2), pages 189–199.

Jungmaier, J., Kassner, N., and Roth, B. (2020). Dirichlet-smoothed word embeddings for low-resource settings. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3560–3565, Marseille, France.

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.

Loftsson, H. and Rögnvaldsson, E. (2007). IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Proceedings of InterSpeech 2007, Special session: "Speech and language technology for less-resourced languages"*, Antwerp, Belgium.

Loftsson, H. and Östling, R. (2013). Tagging a Morphologically Complex Language Using an Averaged Perceptron Tagger: The Case of Icelandic. In *Proceedings of the 19th Nordic conference of computational linguistics (NODALIDA-2013), NEALT Proceedings Series 16*, Oslo, Norway.

Magnússon, B. (2019). Ég gaf ambáttina konunginum. Umröðun tveggja andlaga í íslensku. BA thesis, University of Iceland.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia.

Mollberg, D. E., Jónsson, Ó. H., Þorsteinsdóttir, S., Steingrímsson, S., Magnúsdóttir, E. H., and Gudnason, J. (2020). Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3463–3467, Marseille, France.

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.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., van den Driessche, G., Hendricks, L. A., Rauh, M., Huang, P.-S., Glaese, A., Welbl, J., Dathathri, S., Huang, S., Uesato, J., Mellor, J. F. J., Higgins, I., Creswell, A., McAleese, N., Wu, A., Elsen, E., Jayakumar, S. M., Buchatskaya, E., Budden, D., Sutherland, E., Simonyan, K., Paganini, M., Sifre, L., Martens, L., Li, X. L., Kuncoro, A., Nematzadeh, A., Gribovskaya, E., Donato, D., Lazaridou, A., Mensch, A., Lespiau, J.-B., Tsimpoukelli, M., Grigorev, N. K., Fritz, D., Sottiaux, T., Pajarskas, M., Pohlen, T., Gong, Z., Toyama, D., de Masson d'Autume, C., Li, Y., Terzi, T., Mikulik, V., Babuschkin, I., Clark, A., de Las Casas, D., Guy, A., Jones, C., Bradbury, J., Johnson, M. G., Hechtman, B. A., Weidinger, L., Gabriel, I., Isaac, W. S., Lockhart, E., Osindero, S., Rimell, L., Dyer, C., Vinyals, O., Ayoub, K. W., Stanway, J., Bennett, L. L., Hassabis, D., Kavukcuoglu, K., and Irving, G. (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *ArXiv*, abs/2112.11446.

Sigurðsson, E. F. and Wood, J. (2021). Icelandic Case Syncretism and the Syntax-Morphology Interface. *Working Papers in Scandinavian Syntax*, 105:18–44.

Sinclair, J. (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. Collins ELT.

Snorrason, O. (2021). Samfall og misræmi í þolmynd: Áhrif samfalls á fall- og samræmiskröfur í íslenskri þolmynd. Master's thesis, University of Iceland.

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)*, pages 4361–4366, 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 on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Bulgaria.

Steingrímsson, S., Barkarson, S., and Örnólfsson, G. T. (2020). Facilitating Corpus Usage: Making Icelandic Corpora More Accessible for Researchers and Language Users. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3399–3405, Marseille, France.

Strømberg-Derczynski, L., Ciosici, M., Baglini, R., Chris-

tiansen, M. H., Dalsgaard, J. A., Fusaroli, R., Henrichsen, P. J., Hvingelby, R., Kirkedal, A., Kjeldsen, A. S., Ladefoged, C., Nielsen, F. Å., Madsen, J., Petersen, M. L., Rystrøm, J. H., and Varab, D. (2021). The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online).

11. Language Resource References

Arnardóttir, Þórunn and Ingason, Anton Karl and Steingrímsson, Steinþór and Helgadóttir, Sigrún and Rögnvaldsson, Eiríkur and Barkarson, Starkaður and Guðnason, Jón. (2020). *The Icelandic Contemporary Treebank (IceConTree)*.

Barkarson, Starkaður and Steingrímsson, Steinþór. (2021a). *IGC-News1-21.05 (The Icelandic Gigaword Corpus: News 1)*.

Barkarson, Starkaður and Steingrímsson, Steinþór. (2021b). *IGC-News2-21.05 (The Icelandic Gigaword Corpus: News 2)*.

Barkarson, Starkaður and Steingrímsson, Steinþór. (2021c). *IGC-Parla-21.05 (The Icelandic Gigaword Corpus: Parliamentary speeches)*.

Barkarson, Starkaður and Sigurðsson, Einar Freyr and Rögnvaldsson, Eiríkur and Hafsteinsdóttir, Hildur and Loftsson, Hrafn and Steingrímsson, Steinþór and Andrésdóttir, Þórdís Dröfn. (2020a). *MIM-GOLD 20.05*.

Barkarson, Starkaður and Steingrímsson, Steinþór and Andrésdóttir, Þórdís Dröfn and Hafsteinsdóttir, Hildur. (2020b). *IGC - evaluation set 20.09*.

Barkarson, Starkaður and Steingrímsson, Steinþór and Daníelsson, Hjalti. (2021a). *IGC-Social-21.10 (The Icelandic Gigaword Corpus - Social media)*.

Barkarson, Starkaður and Steingrímsson, Steinþór and Hafsteinsdóttir, Hildur and Andrésdóttir, Þórdís Dröfn and Eiríksdóttir, Inga Guðrún and Magnússon, Bolli. (2021b). *IGC-Journals-21.12 (The Icelandic Gigaword Corpus - scholarly and scientific journals)*.

Barkarson, Starkaður and Steingrímsson, Steinþór and Hafsteinsdóttir, Hildur and Ingimundarson, Finnur. (2021c). *IGC-Books-21.10*.

David Graff, Christopher Cieri. (2003). *English Gigaword LDC2003T05*. Philadelphia: Linguistic Data Consortium.

Erjavec, Tomaž and Ogrodniczuk, Maciej and Osenova, Petya and Ljubešić, Nikola and Simov, Kiril and Grigorova, Vladislava and Rudolf, Michał and Pančur, Andrej and Kopp, Matyáš and Barkarson, Starkaður and Steingrímsson, Steinþór and van der Pol, Henk and Depoorter, Griet and de Does, Jesse and Jongejan, Bart and Haltrup Hansen, Dorte and Navarretta, Costanza and Calzada Pérez, María and de Macedo, Luciana D. and van Heusden, Ruben and Marx, Maarten and Çöltekin, Çağrı and Coole, Matthew and Agnoloni, Tommaso and Frontini, Francesca and Montemagni, Simonetta and Quochi, Valeria and Venturi, Giulia and Ruisi, Manuela and Marchetti, Carlo and Battistoni, Roberto and Sebők, Miklós and Ring, Orsolya and Darģis, Roberts and Utka, Andrius and Petkevičius, Mindaugas and Briedienė, Monika and Krilavičius, Tomas and Morkevičius, Vaidas and Diwersy, Sascha and Luxardo, Giancarlo and Rayson, Paul. (2021). *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*.

Friðriksdóttir, Steinunn Rut and Daníelsson, Hjalti and Steingrímsson, Steinþór. (2021). *IceBATS – The Icelandic Bigger Analogy Test Set.* CLARIN-IS, http://hdl.handle.net/20.500.12537/120.

Friðriksdóttir, Steinunn Rut and Daníelsson, Hjalti and Steingrímsson, Steinþór. (2022a). *Word Embeddings - FastText optimized for IceBATS 22.04.* CLARIN-IS, http://hdl.handle.net/20.500.12537/211.

Friðriksdóttir, Steinunn Rut and Daníelsson, Hjalti and Steingrímsson, Steinþór. (2022b). *Word Embeddings - GloVe optimized for IceBATS 22.04.* CLARIN-IS, http://hdl.handle.net/20.500.12537/210.

Friðriksdóttir, Steinunn Rut and Daníelsson, Hjalti and Steingrímsson, Steinþór. (2022c). *Word Embeddings - Word2Vec optimized for IceBATS 22.04.* CLARIN-IS, http://hdl.handle.net/20.500.12537/209.

Helgadóttir, Sigrún and Barkarson, Starkaður and Hafsteinsdóttir, Hildur and Andrésdóttir, Þórdís Dröfn. (2020). *Icelandic Frequency Dictionary 2020.05 - training/testing sets*.

Helgadóttir, Sigrún. (2018). *Icelandic Frequency Dictionary 2018.10 - training/testing sets*.

Ingólfsdóttir, Svanhvít Lilja and Guðjónsson, Ásmundur Alma and Loftsson, Hrafn. (2020). *MIM-GOLD-NER – named entity recognition corpus (2021-09-29)*.

Loftsson, Hrafn and Yngvason, Jökull H. and Helgadóttir, Sigrún and Rögnvaldsson, Eiríkur and Barkarson, Starkaður. (2018). *MIM-GOLD 1.0*.

Símonarson, Haukur Barri and Snæbjarnarson, Vésteinn and Þorsteinsson, Vilhjálmur. (2020). *En-Is Synthetic Parallel Corpus*.

Steingrímsson, Steinþór and Barkarson, Starkaður. (2021). *IGC-Adjud-21.05 (The Icelandic Gigaword Corpus: Adjudications)*.

Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell and Ackermann, Elia and Aepli, Noëmi and Aghaei, Hamid and Agić, Željko and Ahmadi, Amir and Ahrenberg, Lars and Ajede, Chika Kennedy and Aleksandravičiūtė, Gabrielė and Alfina, Ika and Antonsen, Lene and Aplonova, Katya and Aquino, Angelina and Aragon, Carolina and Aranzabe, Maria Jesus and Arıcan, Bilge Nas and Arnardóttir, Ĥórunn and Arutie, Gashaw and Arwidarasti, Jessica Naraiswari and Asahara, Masayuki and Aslan, Deniz Baran and Ateyah, Luma and Atmaca, Furkan and Attia, Mohammed and Atutxa, Aitziber and Augustinus, Liesbeth and Badmaeva, Elena and Balasubramani, Keerthana and Ballesteros, Miguel and Banerjee, Esha and Bank, Sebastian and Barbu Mititelu, Verginica and Barkarson, Starkaður and Basile, Rodolfo and Basmov, Victoria and Batchelor, Colin and Bauer, John and Bedir, Seyyit Talha and Bengoetxea, Kepa and Berk, Gözde and Berzak, Yevgeni and Bhat, Irshad Ahmad and Bhat, Riyaz Ahmad and Biagetti, Erica and Bick, Eckhard and Bielinskienė, Agnė and Bjarnadóttir, Kristín and Blokland, Rogier and Bobicev, Victoria and Boizou, Loïc and Borges Völker, Emanuel and Börstell,

Carl and Bosco, Cristina and Bouma, Gosse and Bowman, Sam and Boyd, Adriane and Braggaar, Anouck and Brokaitė, Kristina and Burchardt, Aljoscha and Candito, Marie and Caron, Bernard and Caron, Gauthier and Cassidy, Lauren and Cavalcanti, Tatiana and Cebiroğlu Eryiğit, Gülşen and Cecchini, Flavio Massimiliano and Celano, Giuseppe G. A. and Čéplö, Slavomír and Cesur, Neslihan and Cetin, Savas and Çetinoğlu, Özlem and Chalub, Fabricio and Chauhan, Shweta and Chi, Ethan and Chika, Taishi and Cho, Yongseok and Choi, Jinho and Chun, Jayeol and Chung, Juyeon and Cignarella, Alessandra T. and Cinková, Silvie and Collomb, Aurélie and Çöltekin, Çağrı and Connor, Miriam and Courtin, Marine and Cristescu, Mihaela and Daniel, Philemon and Davidson, Elizabeth and de Marneffe, Marie-Catherine and de Paiva, Valeria and Derin, Mehmet Oguz and de Souza, Elvis and Diaz de Ilarraza, Arantza and Dickerson, Carly and Dinakaramani, Arawinda and Di Nuovo, Elisa and Dione, Bamba and Dirix, Peter and Dobrovoljc, Kaja and Dozat, Timothy and Droganova, Kira and Dwivedi, Puneet and Eckhoff, Hanne and Eiche, Sandra and Eli, Marhaba and Elkahky, Ali and Ephrem, Binyam and Erina, Olga and Erjavec, Tomaž and Etienne, Aline and Evelyn, Wograine and Facundes, Sidney and Farkas, Richárd and Ferdaousi, Jannatul and Fernanda, Marília and Fernandez Alcalde, Hector and Foster, Jennifer and Freitas, Cláudia and Fujita, Kazunori and Gajdošová, Katarína and Galbraith, Daniel and Garcia, Marcos and Gärdenfors, Moa and Garza, Sebastian and Gerardi, Fabrício Ferraz and Gerdes, Kim and Ginter, Filip and Godoy, Gustavo and Goenaga, Iakes and Gojenola, Koldo and Gökırmak, Memduh and Goldberg, Yoav and Gómez Guinovart, Xavier and González Saavedra, Berta and Griciūtė, Bernadeta and Grioni, Matias and Grobol, Loïc and Grūzītis, Normunds and Guillaume, Bruno and Guillot-Barbance, Céline and Güngör, Tunga and Habash, Nizar and Hafsteinsson, Hinrik and Hajič, Jan and Hajič jr., Jan and Hämäläinen, Mika and Hà Mỹ, Linh and Han, Na-Rae and Hanifmuti, Muhammad Yudistira and Hardwick, Sam and Harris, Kim and Haug, Dag and Heinecke, Johannes and Hellwig, Oliver and Hennig, Felix and Hladká, Barbora and Hlaváčová, Jaroslava and Hociung, Florinel and Hohle, Petter and Huber, Eva and Hwang, Jena and Ikeda, Takumi and Ingason, Anton Karl and Ion, Radu and Irimia, Elena and Ishola, Ọlájídé and Ito, Kaoru and Jannat, Siratun and Jelínek, Tomáš and Jha, Apoorva and Johannsen, Anders and Jónsdóttir, Hildur and Jørgensen, Fredrik and Juutinen, Markus and K, Sarveswaran and Kaşıkara, Hüner and Kaasen, Andre and Kabaeva, Nadezhda and Kahane, Sylvain and Kanayama, Hiroshi and Kanerva, Jenna and Kara, Neslihan and Katz, Boris and Kayadelen, Tolga and Kenney, Jessica and Kettnerová, Václava and Kirchner, Jesse and Klementieva, Elena and Klyachko, Elena and Köhn, Arne and Köksal, Abdullatif and Kopacewicz, Kamil and Korkiakangas, Timo and Köse, Mehmet and Kotsyba, Natalia and Kovalevskaitė, Jolanta and Krek, Simon and Krishnamurthy, Parameswari and Kübler, Sandra and Kuyrukçu, Oğuzhan and Kuzgun, Aslı and Kwak, Sookyoung and Laippala, Veronika and Lam, Lucia and Lambertino, Lorenzo and Lando, Tatiana and Larasati, Septina Dian and Lavrentiev, Alexei and Lee, John and Lê Hồng, Phuong and Lenci, Alessandro and Lertpradit, Saran and Leung, Herman and Levina, Maria and Li, Cheuk Ying and Li, Josie and Li, Keying and Li, Yuan and Lim, KyungTae and Lima Padovani, Bruna and Lindén, Krister and Ljubešić, Nikola and Loginova, Olga and Lusito, Stefano and Luthfi, Andry and Luukko, Mikko and Lyashevskaya, Olga and Lynn, Teresa and Macketanz, Vivien and Mahamdi, Menel and Maillard, Jean and Makazhanov, Aibek and Mandl, Michael and Manning, Christopher and Manurung, Ruli and Marşan, Büşra and Mărănduc, Cătălina and Mareček, David and Marheinecke, Katrin and Martínez Alonso, Héctor and Martín-Rodríguez, Lorena and Martins, André and Mašek, Jan and Matsuda, Hiroshi and Matsumoto, Yuji and Mazzei, Alessandro and McDonald, Ryan and McGuinness, Sarah and Mendonça, Gustavo and Merzhevich, Tatiana and Miekka, Niko and Mischenkova, Karina and Misirpashayeva, Margarita and Missilä, Anna and Mititelu, Cătălin and Mitrofan, Maria and Miyao, Yusuke and Mojiri Foroushani, AmirHossein and Molnár, Judit and Moloodi, Amirsaeid and Montemagni, Simonetta and More, Amir and Moreno Romero, Laura and Moretti, Giovanni and Mori, Keiko Sophie and Mori, Shinsuke and Morioka, Tomohiko and Moro, Shigeki and Mortensen, Bjartur and Moskalevskyi, Bohdan and Muischnek, Kadri and Munro, Robert and Murawaki, Yugo and Müürisep, Kaili and Nainwani, Pinkey and Nakhlé, Mariam and Navarro Horñiacek, Juan Ignacio and Nedoluzhko, Anna and Nešpore-Bērzkalne, Gunta and Nevaci, Manuela and Nguyễn Thị, Luong and Nguyễn Thị Minh, Huyền and Nikaido, Yoshihiro and Nikolaev, Vitaly and Nitisaroj, Rattima and Nourian, Alireza and Nurmi, Hanna and Ojala, Stina and Ojha, Atul Kr. and Olúókun, Adéday' and Omura, Mai and Onwuegbuzia, Emeka and Osenova, Petya and Östling, Robert and Øvrelid, Lilja and Özateş, Şaziye Betül and Özçelik, Merve and Özgür, Arzucan and Öztürk Başaran, Balkız and Park, Hyunji Hayley and Partanen, Niko and Pascual, Elena and Passarotti, Marco and Patejuk, Agnieszka and Paulino-Passos, Guilherme and Peljak-Łapińska, Angelika and Peng, Siyao and Perez, Cenel-Augusto and Perkova, Natalia and Perrier, Guy and Petrov, Slav and Petrova, Daria and Phelan, Jason and Piitulainen, Jussi and Pirinen, Tommi A and Pitler, Emily and Plank, Barbara and Poibeau, Thierry and Ponomareva, Larisa and Popel, Martin and Pretkalniņa, Lauma and Prévost, Sophie and Prokopidis, Prokopis and Przepiórkowski, Adam and Puolakainen, Tiina and Pyysalo, Sampo and Qi, Peng and Rääbis, Andriela and Rademaker, Alexandre and Rahoman, Mizanur and Rama, Taraka and Ramasamy, Loganathan and Ramisch, Carlos and Rashel, Fam and Rasooli, Mohammad Sadegh and Ravishankar, Vinit and Real, Livy and Rebeja, Petru and Reddy, Siva and Regnault, Mathilde and Rehm, Georg and Riabov, Ivan and Rießler, Michael and Rimkutė, Erika and Rinaldi, Larissa and Rituma, Laura and Rizqiyah, Putri and Rocha, Luisa and Rögnvaldsson, Eiríkur and Romanenko, Mykhailo and Rosa,

Rudolf and Roșca, Valentin and Rovati, Davide and Rudina, Olga and Rueter, Jack and Rúnarsson, Kristján and Sadde, Shoval and Safari, Pegah and Sagot, Benoît and Sahala, Aleksi and Saleh, Shadi and Salomoni, Alessio and Samardžić, Tanja and Samson, Stephanie and Sanguinetti, Manuela and Sanıyar, Ezgi and Särg, Dage and Saulīte, Baiba and Sawanakunanon, Yanin and Saxena, Shefali and Scannell, Kevin and Scarlata, Salvatore and Schneider, Nathan and Schuster, Sebastian and Schwartz, Lane and Seddah, Djamé and Seeker, Wolfgang and Seraji, Mojgan and Shahzadi, Syeda and Shen, Mo and Shimada, Atsuko and Shirasu, Hiroyuki and Shishkina, Yana and Shohibussirri, Muh and Sichinava, Dmitry and Siewert, Janine and Sigurðsson, Einar Freyr and Silveira, Aline and Silveira, Natalia and Simi, Maria and Simionescu, Radu and Simkó, Katalin and Šimková, Mária and Simov, Kiril and Skachedubova, Maria and Smith, Aaron and Soares-Bastos, Isabela and Sourov, Shafi and Spadine, Carolyn and Sprugnoli, Rachele and Steingrímsson, Steinþór and Stella, Antonio and Straka, Milan and Strickland, Emmett and Strnadová, Jana and Suhr, Alane and Sulestio, Yogi Lesmana and Sulubacak, Umut and Suzuki, Shingo and Szántó, Zsolt and Taguchi, Chihiro and Taji, Dima and Takahashi, Yuta and Tamburini, Fabio and Tan, Mary Ann C. and Tanaka, Takaaki and Tanaya, Dipta and Tella, Samson and Tellier, Isabelle and Testori, Marinella and Thomas, Guillaume and Torga, Liisi and Toska, Marsida and Trosterud, Trond and Trukhina, Anna and Tsarfaty, Reut and Türk, Utku and Tyers, Francis and Uematsu, Sumire and Untilov, Roman and Urešová, Zdeňka and Uria, Larraitz and Uszkoreit, Hans and Utka, Andrius and Vajjala, Sowmya and van der Goot, Rob and Vanhove, Martine and van Niekerk, Daniel and van Noord, Gertjan and Varga, Viktor and Villemonte de la Clergerie, Eric and Vincze, Veronika and Vlasova, Natalia and Wakasa, Aya and Wallenberg, Joel C. and Wallin, Lars and Walsh, Abigail and Wang, Jing Xian and Washington, Jonathan North and Wendt, Maximilan and Widmer, Paul and Wijono, Sri Hartati and Williams, Seyi and Wirén, Mats and Wittern, Christian and Woldemariam, Tsegay and Wong, Tak-sum and Wróblewska, Alina and Yako, Mary and Yamashita, Kayo and Yamazaki, Naoki and Yan, Chunxiao and Yasuoka, Koichi and Yavrumyan, Marat M. and Yenice, Arife Betül and Yıldız, Olcay Taner and Yu, Zhuoran and Yuliawati, Arlisa and Žabokrtský, Zdeněk and Zahra, Shorouq and Zeldes, Amir and Zhou, He and Zhu, Hanzhi and Zhuravleva, Anna and Ziane, Rayan. (2021). *Universal Dependencies 2.9*.