ALEXIA: A Lexicon Acquisition Tool

Steinunn Rut Friðriksdóttir, Atli Jasonarson,
Steinþór Steingrímsson, and Einar Freyr Sigurðsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
{srf2, atj9}@hi.is,
{steinthor.steingrimsson, einar.freyr.sigurdsson}@arnastofnun.is

Abstract

We present a new corpus tool, ALEXIA, which is designed to facilitate research using the Icelandic Gigaword Corpus but can be adapted to any text corpus. The tool aids the compilation and expansion of lexical databases and dictionaries by comparing the vocabulary of the database to that of the corpus in order to find gaps in the data. In particular, two well-known Icelandic language resources are incorporated into the design in order to explore the tool's usage. We describe the design and functionality of the tool, how it can be adapted to various data sources and the process of filtering out noise in order to get a clean list of word candidates. Additionally, we present an extensive list of manually collected stop words that can be used to minimize distortion in research results.

1 Introduction

Using automated methods can expedite the process of compiling dictionaries and lexical resources by identifying and collecting candidate words not included in the lexicon from a relevant text corpus. The vocabulary of the corpus is examined in order to find information that is representative of word frequencies in a given language or domain. While manual analysis has certainly made an important contribution throughout the centuries, the availability of computers radically changed lexicography in the middle of the 20th century (Kennedy, 1998, 5), when corpus linguistics shifted the focus towards a quantitative and descriptive approach to language analysis (Bonelli, 2010).

In this paper, we present a new corpus tool, ALEXIA (Friðriksdóttir and Jasonarson, 2021), whose purpose is to facilitate research in lexicography using the Icelandic Gigaword Corpus (IGC). In Section 2, we discuss the motivation behind the tool and its relation to the CLARIN infrastructure. Section 3 briefly discusses previous work and describes the aforementioned corpus. Section 4 describes the design and architecture of the tool, the filtering applied in order to get the best possible results and our manual compilation of corpus-specific stop words. We conclude in Section 5.

2 Motivation and Relation to CLARIN Infrastructure

In recent years, language technology (LT) has gained some momentum in Iceland, most recently in connection to a national language technology program aimed at building basic LT resources (Nikulásdóttir et al., 2020). This has resulted in a considerable increase of publicly available language resources and LT tools. All data created within the program are distributed with open licenses and made universally and permanently available in the CLARIN-IS repository. ALEXIA will serve as a useful tool to compile larger and better lexical resources within the program and after it comes to an end, and thus serve to improve the data available within the CLARIN infrastructure.

As the number of publicly available Icelandic language resources increases, the creation of lexical databases becomes easier. With the growing volume of data, however, the task of manually overseeing potential data gaps becomes ever more cumbersome. ALEXIA is intended to facilitate the construction or expansion of these lexical databases by sourcing large data sets in order to detect potential gaps in their

vocabulary. ALEXIA offers a structured way for the editors of these databases to compare the vocabulary to that of large corpora and at the same time collect statistics on the words' frequencies, individual word form frequencies and frequencies within specific types of text. The tool can also be used to compose new dictionaries, specialized dictionaries, terminologies or other lexicons.

ALEXIA also offers academic research potential, not least in a sociolinguistic context. When new words gain foothold in the language or when a word gains popularity in an unusual context, it can be an indicator of societal changes (Michel et al., 2011). To name examples from different areas of society, advances in technology require an entirely new vocabulary and gender equality movements and progression in LBTQ+ rights have been making their influence in heavily gendered languages and demand an updated vocabulary. A more recent example is the skyrocketing use of pandemic-related vocabulary that goes hand in hand with changes in the development of COVID-19, see e.g. Þorbergsdóttir and Steingrímsson (2021). Examining changes in word frequencies in various data or from different years is made easy by the use of ALEXIA.

3 Previous Work and Relation to the Icelandic Gigaword Corpus

Corpus-driven approaches to lexicography gained prominence in the 1980s, not least due to the success of the COBUILD project (Sinclair, 1987). Corpus-linguistic tools are often dependent on specific corpora, such as the Danish KorpusDK,[1] but can also be generic and corpus-independent, an example of which is XAIRA, an XML-based tool created for but not limited to the British National Corpus (Xiao, 2006). The functionalities of such tools can vary from simple frequency counts, as can be seen in Michael Barlow's MonoConcEsy,[2] to advanced features such as can be seen in Sketch Engine (Kilgarriff et al., 2014). ALEXIA is designed for the Icelandic LT-resource environment, centering around the IGC which is distributed in TEI P5 format, with options to take advantage of common lexical resources imported as plain text files. The tool can easily be adapted to other languages, having similar resource environments.

ALEXIA is designed to be used with the IGC (Steingrímsson et al., 2018) but is not limited to it. The IGC is a corpus of mostly contemporary texts in continuous collection, the 2019 version consisting of around 1.6 billion running words of text along with their morphosyntactic tags and lemmas. It is by far the largest available text corpus in Icelandic and therefore serves well as a representation of actual usage of the Icelandic language. It has various subcorpora from a variety of genres including parliamentary speeches, literature, media texts, etc. One of the primary objectives of the compilation of the IGC is that it is openly available and constantly expanding. Thus, it is ideal for experimenting with a tool like ALEXIA.

4 Architecture

In this section we discuss ALEXIA's design and intended use. As stated in Section 3, the tool is designed for but not limited to the IGC. ALEXIA is run through the command line and has two language options, Icelandic and English. The user can select the default settings of the tool, which involves either the Database of Icelandic Morphology (Bjarnadóttir et al., 2019) or the Dictionary of Contemporary Icelandic (Jónsdóttir and Úlfarsdóttir, 2019), or they can proceed to choose their own input data. In either case, the user is guided towards setting up the appropriate databases for the lexicon to be examined. Additionally, if the default settings are chosen, a filter database containing approximately 60 thousand stop words is created (discussed in Section 4.2). If the user chooses to use their own data, a filter database can also be created from a list of words provided by the user. The stop words can easily be expanded or modified depending on the individual user's need.

4.1 The Databases

The Database of Icelandic Morphology (DIM) contains inflectional paradigms of 303,067 words as of September 2021. Its development has been continuous since 2004 and its applicability for various assignments has increased over the years. It is widely used not only as a linguistic resource in academic

[1] https://ordnet.dk/korpusdk
[2] https://www.monoconc.com/

research and the development of LT resources but also as a reference for the general public. Dictionary of Contemporary Icelandic (DCI) was first published in 2016 and has been in constant expansion since then. It is based on the multi-lingual Scandinavian dictionary ISLEX (Úlfarsdóttir, 2014) and includes various information such as pronunciation audio, illustrative images, collocations and fixed expressions. As these two databases are in constant expansion, there is a need for quick and concise ways to determine gaps in the data. These databases are therefore ideal for exploring the functionalities of ALEXIA.

As previously stated, the databases are compared to the IGC. The corpus can be used in its entirety, representing the most common registers of written Icelandic, or specific subcorpora can be chosen. The latter option can be beneficial if the user intends to research a specific vocabulary (e.g. sports).

4.2 Filters and Stop Words

ALEXIA is designed around the IGC and therefore its default settings have a number of predefined rules for filtering a word candidate list. Using a tagged corpus can be very beneficial as it provides the option to exclude words tagged with certain parts of speech from the results that may not be well suited for dictionary compilations. Lemmatization also provides the option to exclude all oblique conjugations in order to lessen potential noise in the results. Words with certain POS-tags are excluded from the results, e.g. emails and websites, abbreviations and exclamations. Additional filters are applied, such as excluding words that start or end in a hyphen, and the user can optionally exclude proper names as they tend to overflood candidate lists.

However, not all errors are caught by these filters. As available corpora have increased in volume, the use of automatic POS-tagging and lemmatization has also increased. While their use has certainly sped up and facilitated work in computational linguistics, it is not without its limitations. When a corpus is not corrected by human annotators there will be a certain amount of noise included, especially when words are incorrectly tagged or lemmatized. Perhaps the largest contribution of our work is therefore the list of stop words compiled from the IGC. We have manually collected over 60 thousand typos, misspellings, outdated spellings (e.g. the use of ‘z’, which was replaced by ‘s’ in the official spelling standard in 1973), OCR errors, foreign words, improperly lemmatized words and other non-word tokens. The list greatly minimizes distortion when generating candidate lists.

4.3 Candidate Lists

After comparing the input data to the corpus and filtering the results, ALEXIA creates a candidate list of the user’s choice. The following lists are available:

1. Frequency lists where either all lemmas or all word forms that are not found in the input lexicon are displayed along with their frequencies in the comparison corpus. This can be utilized to estimate the usage of certain words and thus decide if they are suitable candidates. Additionally, information on nouns where the plural form is much more frequent than the singular form can be included, suggesting that a word might only exist in the plural.

2. Frequency lists including the top five collocation examples taken from the corpus. The collocations include the two previous words and the two words following the candidate word. This can be used to determine the context of a word, especially if it is often used in fixed expressions.

3. Frequency lists, where all word forms that appear with a given lemma in the comparison corpus are displayed. This can be useful for determining if a certain word form is the most common and thus if the word is only used in a certain context, e.g. a fixed expression. If word forms are chosen, all lemmas that appear with the word form are displayed. This can be useful for determining if a word can belong to multiple word classes and whether the automatized lemmatization delivers the expected results.

4. Frequency lists where individual word frequencies within certain types of text are displayed. This can be useful for building specialized vocabularies (e.g. related to news, math or sports domains).

5 Conclusion and Future Work

The lexicon acquisition tool, ALEXIA, aims to facilitate the creation and expansion of lexicons and dictionaries by comparing their vocabularies to text corpora, such as the Icelandic Gigaword Corpus. Additionally, it can be used as a resource for academic research, not least concerning sociolinguistics. We include several filtering options in order to deliver appropriate candidate lists, including a stop-word list, a list of POS-tags to exclude and an option to filter out proper nouns if desired. The resulting lists are based on frequencies within the comparison corpus but can include additional information intended to make the lexicographer's work as fast and easy as possible.

As the availability and size of corpora expands, so does the need for a corpus tool like the one we have presented here. Future development of ALEXIA could include ways to visualize in a graph the timeline of a word's use and thus give a better overview of when the word gained foothold in the language or if its usage is declining. We encourage anyone interested to use the tool and modify it as they wish.

Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019–2023. The program, which is managed and coordinated by Almannarómur (https://almannaromur.is/), is funded by the Icelandic Ministry of Education, Science and Culture.

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*, pages 146–154, Turku, Finland.

Bonelli, E. T. 2010. Theoretical overview of the evolution of corpus linguistics. In O'Keeffe, A. and McCarthy, M., editors, *The Routledge Handbook of Corpus Linguistics*, pages 14–28. Routledge, London and New York.

Friðriksdóttir, S. R. and Jasonarson, A. 2021. *ALEXIA: Lexicon Acquisition Tool for Icelandic 3.0*. CLARIN-IS, http://hdl.handle.net/20.500.12537/123.

Jónsdóttir, H. and Úlfarsdóttir, Þ. 2019. Íslensk nútímamálsorðabók. *Orð og tunga*, 21:1–26.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Longman, London and New York.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., and Suchomel, V. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, 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.

Nikulásdóttir, A. B., 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 Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 3414–3422, Marseille, France.

Sinclair, J., editor. 1987. *Looking up: an account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. Collins, London.

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.

Úlfarsdóttir, Þ. 2014. ISLEX – a Multilingual Web Dictionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2820–2825, Reykjavik, Iceland.

Xiao, R. 2006. Xaira – an XML aware indexing and retrieval architecture. *Corpora*, 1(1):99–103.

Þorbergsdóttir, Á. and Steingrímsson, S. 2021. Orð ársins 2020: Sóttkví [Word of the Year 2020]. *Hugrás,* https://hugras.is/2021/01/ord-arsins-2020-sottkvi/.