Natural Language Processing for Everyone: The Case for a Centralized Icelandic NLP Platform

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

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

Abstract

This paper presents the Árni Magnússon Institute's NLP platform, a website where the functionality of important language technology tools for Icelandic is made accessible to the general public, both through a user interface and a standardized API. The platform is presented as a solution to the problem that even though a specific language technology tool exists, it does not go without saying that anyone can use it: A certain technical know-how is almost always a prerequisite for being able to use these tools. The platform is presented with the following NLP solutions integrated: Tokenization, PoS-tagging, Lemmatization and Hyphenation, with more tools becoming available pending future development. The platform is now made available to the public, with the caveats of it being in active development and undergoing regular changes.

Keywords

Icelandic, Language Technology, NLP, Accessibility, API

1. Introduction

Recent years have seen the development and release of many natural language processing (NLP) tools for Icelandic. Thanks to intensive government support within a national language technology programme (Nikulásdóttir, Guðnason, and Steingrímsson 2017; Nikulásdóttir et al. 2020) and the hard work of professionals in the field, these tools have managed to cover most of the basic tasks of language technology (LT) for Icelandic with good results. Most new software and datasets are in open access and have been made publicly available on the Icelandic-language repository of the *Common Language Resources and Technology Infrastructure* (CLARIN-ERIC); CLARIN-IS[1], in addition to other informal open-source projects accessible through standard venues, e.g. GitHub. However, a problem still remains: Even though a specific language technology tool exists, it does not go without saying that anyone can use it 'out of the box'. A certain technical know-how is, to a certain extent, always a prerequisite for being able to use these resources.

DHNB'25: Digital Humanities in the Nordic and Baltic Countries, March 5–7, 2025, Tartu, Estonia
hinrik.hafsteinsson@arnastofnun.is (H. Hafsteinsson); steinthor.steingrimsson@arnastofnun.is (S. Steingrímsson)
https://hinrik.hafsteinsson.is (H. Hafsteinsson); https://steinst.is/ (S. Steingrímsson)
0009-0003-3348-2579 (H. Hafsteinsson); 0000-0002-9776-9507 (S. Steingrímsson)

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

[1]See: https://repository.clarin.is

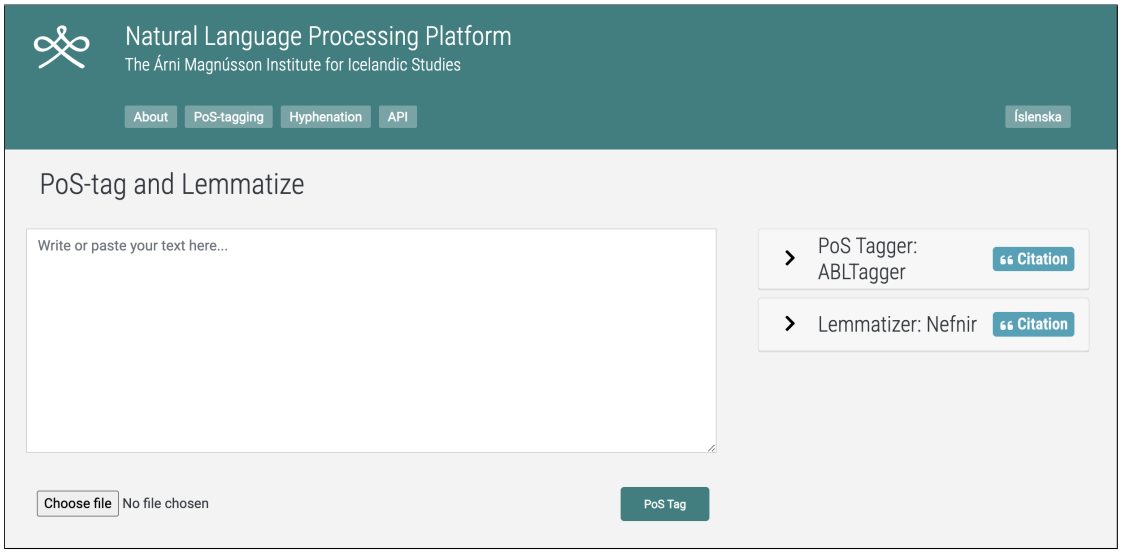

Figure 1: The the PoS-tagging and lemmatization interface of the Árni Magnússon NLP platform.

The Árni Magnússon Institute's NLP platform[2] is presented here as one solution to this problem. The platform is a website where specific language technology tools are made accessible to the general public, both through a user interface and formal APIs. The PoS-tagging and lemmatization interface of the website is shown in Figure 1. The website is now made available to the public, with the caveat of active development.

Below is an outline of the platform's background and development, as well as a discussion of the need for accessible language technology tools in the case of Icelandic. Section 2 gives a brief overview of language technology tools and the current measures towards accessibility in the field. Section 3 presents the Árni Magnússon Institute's NLP platform, its tools, and its API. Section 4 concludes.

2. Language Technology in Iceland

2.1. Language technology tools

Tools in NLP and language technology in general are often classified hierarchically. Certain tasks and tools are 'simple' and serve as the basis for other tools. Examples of these 'simpler' tasks are:

- **Tokenization**: Text is segmented into independent units called tokens. These are most often individual words, punctuation marks, although more specialized tokenization exists.
- **PoS-tagging**: Tokens within text are labelled in a specific way with tags, e.g. by grammatical classification or other linguistic information.

[2]The platform can be accessed at https://malvinnsla.arnastofnun.is/.

- **Lemmatization**: The base forms (lemmas) of words are found, e.g. the word 'running' being lemmatized as 'run'. PoS-tagging the given text is often a prerequisite of this task.
- **Parsing**: The syntax of a given text is analyzed automatically, to produce a syntax 'tree', which encodes the text's sentences' syntactic structure.

It is not uncommon in text processing tasks that use language technology to first tokenize text, then tag it, and finally parse it, with the output of one task being the input of the next.

Other language technology solutions are then considered 'more complex', likely because the task they perform is inherently more complex in itself. Good examples are machine translation, automatic speech recognition, named entity recognition, sentiment analysis, not to mention large language models (LLMs; e.g., ChatGPT etc.). This binary distinction of language technology tools is not exhaustive for all circumstances within language technology, but it proves useful in everyday contexts — not least when the question is: What tools would be good to have easy access to?

2.2. The need for accessible language technology tools

Even though the focus (from the perspective of a software specialist) is on the fact that development in language technology can lead to more language technology, it must not be forgotten that the topic in our field, i.e. the Icelandic language, has a universal reference within Icelandic society. This means that anyone who works with the language, whether in text or speech, may have possible use for language technology tools and solutions. A good example of this is the *Database of Icelandic Morphology* (Bjarnadóttir, Hlynsdóttir, and Steingrímsson 2019) which has proven its applicability in more than one societal domain: It is used by the writing class as a reference for correct inflections and spelling, it is used by Scrabble-players as a reference for allowed words and it is used by language technologists aiming to build applications that adhere to accepted paradigms. Another, more recent example in the same field is the web portal M.is, where machine translation in Icelandic is made accessible, along with dictionaries maintained by the Árni Magnússon Institute.

A fair example of a tool that has significant potential for everyday use, but limited distribution, is a PoS tagger. As mentioned earlier, PoS-tagging consists of labelling words in text with some kind of tag. In the case of Icelandic, words are most commonly labeled with grammatical information: A word's tag could contain information on grammatical gender, number, case, etc. Tagging is the basis of annotated language corpora (e.g. the Icelandic Gigaword Corpus (Steingrímsson et al. 2018; Barkarson, Steingrímsson, and Hafsteinsdóttir 2022) and others) and many PoS-tagging tools have been developed for Icelandic over the years.

Furthermore, as most Icelandic PoS-tagging systems are based on morphological information of words, solutions to more fine-grained tasks can be found in the output of a PoS-tagging tool. An example of this is lexical classification, i.e. analyzing each word in a text into a specific lexical class and potentially even statistical analysis of text based on the content's lexical classes[3].

[3]The usual term for this is a *Part of Speech* (PoS), which gives the term PoS-tagging. This means that in some cases, lexical classificiation and PoS-tagging can be equivalent tasks, as is most often the case for English. This is generally not true for Icelandic, as the mainstream approaches of PoS-tagging Icelandic has lexical category as only a subset of the information encoded in a word's PoS tag. To avoid confusion here, the alternate term *lexical classification* is used for this specific task.

Manual lexical classification has been used by teachers to evaluate the writing of their students and is, for example, one of the tasks of speech-language pathologists, who evaluate the content of their clients' speech. Similarly, this is an important analysis method in linguistic research on Icelandic outside explicit language technology. This means that in these specific examples, automatic PoS-tagging in lexical classification would be very useful.

It is not, however, part of the general, knowledge-base of teachers, speech-language pathologists, and researchers in the humanities to develop, retrieve or run software that is generally used in software development and research. This means that those who would need automatic text PoS-tagging are forced to use manual methods, which can be time-consuming and even inconvenient. In the case of PoS-tagging tools, there is therefore reason to consider whether the tools can be made accessible to those who may need them, regardless of direct technical knowledge. It may be assumed that the same applies to other types of language technology tools. Put plainly, facilitating access to LT tools is facilitating digital humanities in the wider sense.

2.3. Current measures towards accessibility

The Icelandic language technology community has, in general, made significant progress in making language technology tools accessible to the general public. The main venue for this is the Icelandic CLARIN-ERIC registry, CLARIN-IS. CLARIN stands for *Common Language Resources and Technology Infrastructure*, which is a part of the *European Research Infrastructure Consortium* (ERIC), an EU initiative. The CLARIN-IS registry is a repository of language technology tools, datasets, and other resources for Icelandic and, in comparison to more generalized platforms like GitHub, offers stricter archiving guarantees and a more formalized structure. As such, the CLARIN-IS registry is a good example of a centralized repository of language technology tools, which is accessible to the general public. However, CLARIN does not facilitate active development (i.e., fine-grained, decentralized version control), which results in the tools available there being static and not necessarily up-to-date.

In addition to the CLARIN-IS registry, the official arm of the Icelandic Language Technology community (*Samstarf um íslenska máltækni*, SÍM) maintains its own organization profile on GitHub[4], where all deliverables and source code of the government sponsored *Language Technology Program for Icelandic 2019-2024* are made available as GitHub repositories. Furthermore, they include references to respective CLARIN-IS repositories where applicable and HuggingFace repositories[5] for models.

An example of a LT tools being made accessible to the general public in a functional form is the various approaches taken by Miðeind ehf[6], a private company focusing on Icelandic language technology. Their generalized text-analyzer Greynir[7] has been the quintessential syntactical parser for Icelandic for a number of years, as well as providing different methods of data processing free of charge. Miðeind's recent, proprietary platform Málstaður[8] is a high-end

[4]See: https://github.com/icelandic-lt
[5]See: https://huggingface.co/icelandic-lt
[6]See: https://mideind.is
[7]See: https://greynir.is/
[8]See: https://malstadur.mideind.is/

NLP platform for Icelandic, which offers spelling and grammar checking, automatic speech recognitionm and automatic question answering, for a subscription fee.

3. A centralized Icelandic NLP platform

We present a novel natural language processing platform, developed and maintained by the Árni Magnússon Instititute for Icelandic Studies. The platform currently offers a handfull of tools, including a PoS tagger, which are accessible through an intuitive user interface.

3.1. The Árni Magnússon Institute's NLP platform

The Árni Magnússon Institute's NLP platform was originally launched in 2018, primarily to make the PoS-tagging tool ABLTagger (Steingrímsson, Kárason, and Loftsson 2019) accessible online. More tools were added, but the structure was small and maintenance minimal. With the current renewal, the scope and purpose of the platform is being redefined, as well as maintenance plans.

The Árni Magnússon Institute is in a special position when it comes to language technology development for Icelandic. Many other parties, universities and private companies also contribute to the development of language technology solutions nationwide, but formal obligations towards the language are greatest at the Árni Magnússon Institute. In addition, the institute has accumulated a great deal of expertise in language technology, as the institute have been involved in LT projects since LT work took off in Iceland in the early 2000s. Employees of the institute who work directly or indirectly in LT have become numerous and various LT tools have been developed in-house with the corresponding expertise. In addition, the institute maintains a substantial collection of LT-related websites: In addition to the aforementioned DIM (BÍN in Icelandic[9] and M.is, one can mention the Icelandic-Scandinavian dictionary project ISLEX[10] (Úlfarsdóttir 2014), the corpus research platform Málheildavefurinn[11] (Steingrímsson, Barkarson, and Örnólfsson 2020) and the specialized term database of Íðorðabankinn[12] (Þorbergsdóttir 2024), to name but a few. A Language processing platform thus fits well into the repertoire of LT-related services that the Árni Magnússon Institute provides online.

3.2. The tools on the platform

The discussion above underscores that some LT tools have been inaccessible to those who may have use for them due to the technical knowledge required to . Furthermore, such a platform can also be useful for the technically minded as the services offered are also available through API. APIs make the functionality of the website accessible to programs, where it is possible to call a specific endpoint on the website's URL in standardized ways. This means that a user who would previously have had to retrieve a specific tool and install it on their own computer can, as needed (and if the tool in question is available on the website), send service requests to the tool over the internet. The interventions will therefore be smaller and thus the convenience greater.

[9] https://bin.arnastofnun.is
[10] https://islex.arnastofnun.is/
[11] https://malheildir.arnastofnun.is/
[12] https://idord.arnastofnun.is/

With these considerations in mind, the renewed Árni Magnússon Institute's NLP platform is presented with the following tools represented:

- **Tokenization** – Tokenizer (Þorsteinsson, Óladóttir, and Loftsson 2019)
- **PoS-tagging** – ABLTagger 3.0 (POS, Steingrímsson, Kárason, and Loftsson 2019)
- **Lemmatization** – Nefnir (Ingólfsdóttir et al. 2019)
- **Hyphenation** – Skiptir (Rúnarsson 2020)
- In addition to these tools, API documentation[13] is made available, built on the OpenAPI[14] specification.

3.3. Using the platform

To illustrate a use case for the platform, we focus on the tasks of tokenization, PoS-tagging and lemmatization. The user interface (UI) for these tasks is shown in Figure 1. In this task, the generated output is presented together in a single output table. When an Icelandic input sentence, e.g., 'Ég stökk á eftir strætó og veifaði' (eng. *I jumped after the bus and waved*) is supplied to the "PoS-tagging" option in the platform's UI, the platform's expected output is shown in Table 1, localized to English for clarity.

Table 1
Example of tokenization, PoS-tagging, and lemmatization, as presented in the Árni Magnússon Institute's NLP platform' UI

Token	Tag	Lemma	PoS tag description
Ég	fp1en	ég	pronoun, personal pronoun, 1st person, singular, nominative case
stökk	sfg1eþ	stökkva	verb, indicative, active voice, 1st person, singular, preterite
á	aa	á	adverb, does not govern case
eftir	af	eftir	adverb, governs case
strætó	nkeþ	strætó	noun, masculine, singular, dative
og	c	og	conjunction
veifaði	sfg1eþ	veifa	verb, indicative, active voice, 1st person, singular, preterite
.	pl	.	punctuation, end of sentence

Table 1 illustrates the platform's implementation of each of the tokenization, PoS-tagging and lemmatization tasks:

- **Tokenization**: The input sentence is split into tokens, where each token is a word or punctuation mark.
- **PoS-Tagging**: Each token is accompanied by a fine-grained morphological tag, based on the Icelandic Frequency Dictionary tagset (Pind, Magnússon, and Briem 1991), revised for the 21.05 version of the MIM-GOLD corpus (Barkarson et al. 2021). The *Description* column contains an 'expanded' version of the tag, where each constituent letter of the tag string is given a verbose explanation.
- **Lemmatization**: The base form of each token is given in the *Lemma* column.

[13]The API documentation can be accessed at http://malvinnsla.arnastofnun.is/docs
[14]The OpenAPI specification can be accessed at https://swagger.io/specification/.

This version of the Árni Magnússon Institute's NLP platform is now made available to the public, with the caveats of it being in active development and undergoing regular changes.

4. Conclusion and future work

We have presented the Árni Magnússon Institute's NLP platform. The platform is intended to make various Icelandic NLP tools available for the general public. Our intention is to make these tools easier to access and use, and thus hopefully making NLP work easier for a wider range of users than before, facilitating both research in digital humanities and everyday use of the Icelandic language.

Future expansions to the functionality of the platform are planned. As we discuss, various fast and reliable 3rd party solutions to various NLP tasks for Icelandic have been produced and are available under permissive licenses. These include named entity recognition (NER) and various tools for syntactic analysis, which are set to be the main focus of future expansion, within the platform's usage scope and hardware limitations. Updates and revisions to the platform's UI are in discussion and although no plans are present for hardware extensions, they will be implemented as the need arises.

The platform, in its current state, is now made available to the public, with the caveats of it being in active development and undergoing regular changes.

References

Barkarson, Starkaður, Þórdís Dröfn Andrésdóttir, Hildur Hafsteinsdóttir, Árni Davíð Magnússon, Kristján Rúnarsson, Steinþór Steingrímsson, Haukur Páll Jónsson, et al. 2021. *MIM-GOLD 21.05.* CLARIN-IS. http://hdl.handle.net/20.500.12537/113.

Barkarson, Starkaður, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022. "Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus." In *Proceedings of the Language Resources and Evaluation Conference,* 2371–2381. Marseille, France. https://aclanthology.org/2022.lrec-1.254.

Bjarnadóttir, Kristín, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. "DIM: The Database of Icelandic Morphology." In *Proceedings of the 22nd Nordic Conference on Computational Linguistics,* 146–154. Turku, Finland. https://aclanthology.org/W19-6116.

Ingólfsdóttir, Svanhvít Lilja, Hrafn Loftsson, Jón Friðrik Daðason, and Kristín Bjarnadóttir. 2019. "Nefnir: A high accuracy lemmatizer for Icelandic." In *Proceedings of the 22nd Nordic Conference on Computational Linguistics,* 310–315. Turku, Finland. https://aclanthology.org/W19-6133.

Nikulásdóttir, Anna, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. "Language Technology Programme for Icelandic 2019-2023." In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 3414–3422. Marseille, France. https://aclanthology.org/2020.lrec-1.418.

Nikulásdóttir, Anna Björk, Jón Guðnason, and Steinþór Steingrímsson. 2017. *Language Technology for Icelandic. Project Plan.* Reykjavík, Iceland: Icelandic Ministry of Science, Culture / Education. https://rafhladan.is/handle/10802/20054.

Pind, Jörgen, Friðrik Magnússon, and Stefán Briem. 1991. "Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]." *The Institute of Lexicography, University of Iceland, Reykjavik, Iceland.*

Rúnarsson, Kristján. 2020. *Skiptir (20.10).* CLARIN-IS. http://hdl.handle.net/20.500.12537/87.

Steingrímsson, Steinþór, Starkaður Barkarson, and Gunnar Thor Örnólfsson. 2020. "Facilitating Corpus Usage: Making Icelandic Corpora More Accessible for Researchers and Language Users." In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 3399–3405. Marseille, France. https://aclanthology.org/2020.lrec-1.416.

Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 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, Steinþór, Örvar Kárason, and Hrafn Loftsson. 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),* 1161–1168. Varna, Bulgaria. https://aclanthology.org/R19-1133.

Úlfarsdóttir, Þórdís. 2014. "ISLEX — a Multilingual Web Dictionary." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* 2820–2825. Reykjavik, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/pdf/672_Paper.pdf.

Þorbergsdóttir, Ágústa, ed. 2024. *Íðorðabankinn.* https://idord.arnastofnun.is. The Árni Magnússon Institute for Icelandic Studies. Accessed on October 22, 2024.

Þorsteinsson, Vilhjálmur, Hulda Óladóttir, and Hrafn Loftsson. 2019. "A Wide-Coverage Context-Free Grammar for Icelandic and an Accompanying Parsing System." In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019),* 1397–1404. Varna, Bulgaria. https://aclanthology.org/R19-1160.