NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	Automatic Terminology Extraction: New Challenges in Terminology Work in Iceland
Forfatter:	Ágústa Þorbergsdóttir, Atli Jasonarson, Finnur Ágúst Ingimundarson, Einar Freyr Sigurðsson, Steinþór Steingrímsson & Hjalti Daníelsson
Kilde:	Nordiska studier i lexikografi 16, 2023, s. 403–413
URL:	https://tidsskrift.dk/nsil/issue/archive

Automatic Terminology Extraction: New Challenges in Terminology Work in Iceland

Ágústa Þorbergsdóttir, Atli Jasonarson, Finnur Ágúst Ingimundarson, Einar Freyr Sigurðsson, Steinþór Steingrímsson & Hjalti Daníelsson

We present TermPortal, a web-based terminology acquisition and management system to support Icelandic terminology work. We discuss its function and report on how it copes with a selected domain-specific field, namely linguistic terms. Two different automatic extraction methods of building a terminology are explored: The first utilizes a tf-idf (term frequency-inverse document frequency) model and returns tokens with a statistics-based score; if it is higher than a certain threshold, a probable technical term is suggested. The second method makes use of Daðason's BiLSTM Compound Splitter for Icelandic (Kvistur), focusing on the different morphological parts which a given term can consist of. We also try combining these two methods.

KEYWORDS: terms, terminology extraction, domains, Icelandic

1. Introduction

Manual collection of terminology is a very time-consuming task.[1] Terminology work in Iceland has a long history and the work is usually carried out by terminology committees. Over 50 terminology committees of various kinds have existed in Iceland for longer or shorter periods, as regards activity and working methods. Usually, those committees are composed of subject-matter experts, working on a voluntary basis, who have devoted their time to manually creating glossaries within their field. Improving the process of building a new terminology — or finding new words to add to an existing one — and speeding it up is therefore of

1 The work presented in this paper was supported by Rannís (Strategic Research and Development Programme; grant #180017-53011). We thank two anonymous reviewers and the editors for their comments. We also thank Eiríkur Rögnvaldsson whose chapter in his open-access book, *Hljóðkerfi og orðhlutakerfi íslensku*, provided the basis for our case study discussed in sections 3 and 4.

utmost importance and very useful for standardizing vocabulary in specialized fields.

This paper introduces the web-based terminology acquisition and management system TermPortal,[2] which supports Icelandic terminology work. It is intended to facilitate lexicographic work in building terminologies. TermPortal helps us identify terms and see how they are used, which is important, e.g., when defining concepts and the boundaries of LSP (language for special purposes) vs. LGP (language for general purposes).

The paper is structured as follows: Section 2 provides an overview of TermPortal and how texts are processed by automatic term extraction tools. Section 3 presents a case study on linguistic terms to test TermPortal's function and precision while section 4 reports on TermPortal's evaluation. Section 5 presents conclusions of the study and future work.

2. TermPortal

TermPortal is a terminology acquisition and management system. TermPortal consists of two main parts. Firstly, the TermPortal workbench includes an automatic pipeline to extract terminology from media and a web platform where users can create, manage and maintain termbases. Secondly, the automatic term extraction (ATE) system is a central component in the TermPortal workbench but can also be used independently. We looked beyond the traditional methods of manual terminology work and tried to simplify the process of preparing, storing, and sharing terminology glossaries.

Users of the system can upload texts which are then processed by ATE tools, tagging potential terminology candidates for the user to accept or decline. The process is as follows: The user uploads a text file and — optionally — specifies a domain, such as medicine, history or linguistics. TermPortal can use termbases for different fields or run without any support from a termbase. The text is then run through a pipeline, consisting of six steps:

1) The text is tokenized into single-word units, and its punctuation marks are removed.

2 https://termportal.arnastofnun.is/

2) The tokenized text is run through a part-of-speech tagger (with ABL-Tagger as default; see Steingrímsson et al. 2019), returning every token along with its corresponding tag.
3) The tags are used to remove unwanted words, such as foreign ones, proper nouns and numbers, as well as single-character units.
4) The tokens are lemmatized using Nefnir (Ingólfsdóttir et al. 2019), whose accuracy improves substantially when supplied with part-of-speech tags.
5) The lemmatized tokens are run through a tf-idf (term frequency-inverse document frequency) model, trained on roughly 17 million words collected from scholarly and scientific journals and websites (Barkarson et al. 2021).
6) At this stage, there are three options:
    a) The tokens whose tf-idf score is above a given threshold are returned as probable technical terms.
    b) The tokens can be split into their stem structure. If a list of known stems and morphological parts exists for a given field, it can be used to identify tokens as probable terms. If, for example, a stem which is a part of a compound is on such a list, the whole compound may be suggested as a term.
    c) These two methods can be used individually or combined.

The tokens the system identifies as probable technical terms are returned to the user via the web interface. The user is then faced with their text where candidate terms are highlighted, see Figure 1. The user is asked to accept or reject the candidates and their decisions are subsequently stored. Previously accepted terms or terms that are in an available termbase are highlighted in green, while unknown candidate terms are highlighted in blue. The accepted terms are added to the termbase and the rejected ones, highlighted in red, are added to a list of candidates rejected as terms in this domain. The rejected ones are not, however, entirely discarded as they can be used to filter out irrelevant suggestions in future use. The accepted terms are also useful, as different morphological parts can serve to expand the collection and therefore improve the method described in 6b). Additionally, the user can, anywhere in the process,

add their own words, such as ones the system missed or the ones it mis-lemmatized.

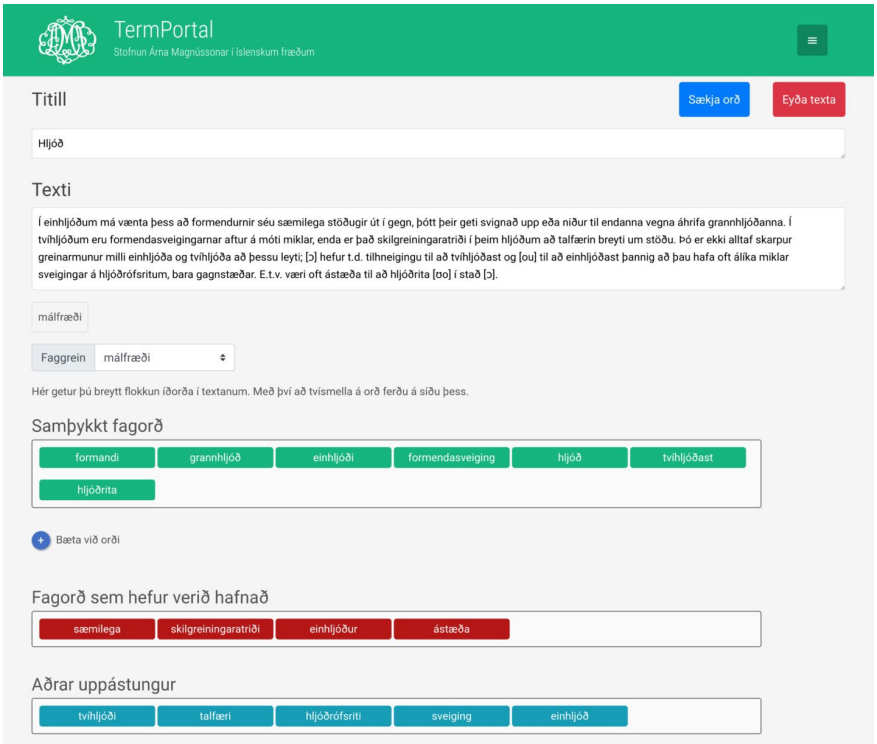

FIGURE 1. A screenshot of TermPortal's web interface in which the user decides which suggestions to accept or reject.

This pipeline — see Figure 2 — only identifies single-word units. We have also experimented with multi-word unit term extraction applying three different methods for identifying term candidates: C-value (Frantzi et al. 2000); an approach based on stem ratio (Daníelsson et al. 2020); and Levenshtein-distance (Levenshtein 1966) between possible candidates and known terms. These approaches are described in more detail in Daníelsson et al. (2020). While these approaches are effective in finding the terms, they also produce many false positives, leading to the effect of the results not being very helpful in boosting productivity of termbase editors. Efforts to raise the accuracy of multi-word term extraction are not within the current scope of our project although we want to study that in future work.

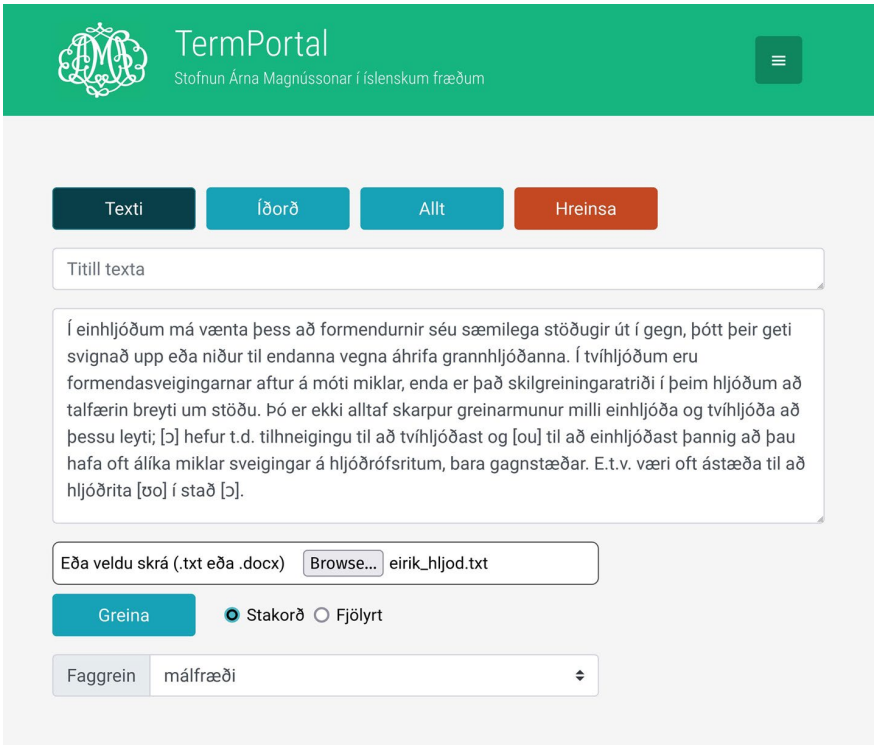

FIGURE 2. A screenshot of the entry point of TermPortal's web interface.

Below the text area in Figure 2 is the analysis button (Icel. *greina*) and two options allowing the user to either extract single-word terms (Icel. *stakorð*) or multi-word terms (Icel. *fjölyrt*), with the latter option still being at an experimental stage. At the bottom, the domain (Icel. *faggrein*) is specified, in this case grammar (Icel. *málfræði*).

3. A case study on linguistic terms

To test and evaluate TermPortal's function and precision, applying the methods discussed in the previous section, we used a chapter on Icelandic speech sounds (chapter 4) from Rögnvaldsson's (2013) book on Icelandic phonology and morphology. The chapter is 16 pages long and it took a single person in the project two hours to mark linguistic terms in the text. The manual registration of terms in the text resulted in 99 terms, which we use as a gold standard (see discussion on evaluation in section 4 below).

One particularity of texts on phonology such as the one chosen for this paper is that some of the terms can be lined up together to form another multi-word term, which raises the question whether that new term should be considered as a sum of its parts and therefore a unique term. A case in point could for example be the classification of vowels in Icelandic, which are divided into either front or back, high or low, or rounded or unrounded vowels. They can furthermore either be long or short. A combination of these terms could therefore be used to describe a certain vowel, for example the sound [i] which is unrounded and the most front and highest vowel in Icelandic, i.e., *frammælt*, *nálægt*, *ókringt sérhljóð* 'front, high, unrounded vowel'. In the present data each adjective was defined and marked individually and combinations such as the example shown here are not considered as terms on their own.

Another classic problem of defining the boundaries of LSP (language for special purposes) vs. LGP (language for general purposes) or simply terminology vs. general vocabulary concerns the use of ordinary adjectives, nouns, etc. as specific terms, as for example the adjective *nálægur*, which has the general meaning 'near(by), close', whereas in the preceding paragraph it has a specific function as a term to describe a certain vowel quality (Eng. *high*). One could also point to the distinction of terms relating to the speech organs, such as between the very much ordinary nouns *tongue* and *lips* as opposed to *palate*, *dorsum linguae* and *uvula*.

4. Evaluation

To evaluate TermPortal's suggestions, the list of 99 terms acts as a 'ground truth' or gold standard, meaning that a perfect model would extract all the 99 terms, and nothing else. That would yield an F_1 score of 1 (or 100%) which is computed by calculating the following:

- True positives (TP): All the terms extracted from the text that are present in the ground truth
- False positives (FP): All the terms extracted from the text that are not present in the ground truth
- False negatives (FN): All the terms not extracted from the text that are present in the ground truth

When these numbers have been calculated, the recall (R) is computed by the equation $\frac{TP}{TP+FN}$, i.e., true positives divided by all terms present in the ground truth. On its own, recall is not very useful, though, because a model could achieve recall of 100% by simply returning all the possible elements. Therefore, precision (P) is used to calculate how accurate the output of the model is: $\frac{TP}{TP+FP}$. The F_1 score is the harmonic mean of these two, which gives us a single number to represent how effective a classifier, such as the one in question, is. It is computed as follows: $2 \times \frac{P \times R}{P+R}$.

As mentioned in section 2, two different automatic methods were taken into consideration and compared to see which one proved to be more effective as a classifier for TermPortal. The first one was based on a tf-idf (term frequency-inverse document frequency) method, running the lemmatized tokens from the text through a tf-idf model specially trained on scholarly and scientific material, returning tokens with a score higher than a certain threshold as probable technical terms, i.e., essentially words considered 'rare' in the model. The output was a diverse list which included amongst other things all or most of the phonetically transcribed words appearing in the text, such as [kʰaltʏr] for *kaldur* 'cold', i.e., not words to be considered as terms. As a result, the efficiency, i.e., the F_1 score, was accordingly rather low and the numbers read as shown in Table 1.

TABLE 1. A method based on tf-idf

True positives:	46
False positives:	65
False negatives:	53
Precision:	41.4%
Recall:	46.5%
F_1:	0.438

The other approach was based on a BiLSTM Compound Splitter for Icelandic (Kvistur) (Daðason et al. 2020) which was used on linguistic terminology from *The Icelandic Term Bank* (https://idord.arnastofnun.is/),

1,367 terms in all. Compounds are extremely common in Icelandic[3] and that also goes for terms where different morphological parts, stems in particular, serve as essential building blocks. As an example, we can take the Icelandic name of the term bank, *Íðorðabankinn*, which we would expect Kvistur to split into the following parts: *íð* 'work, profession', *orða* 'words' and *bankinn* '(the) bank'.

The resulting list, i.e., from using the compound splitter on linguistic terminology, contained several incorrectly split parts, but these were relatively few. One such example were the parts *for* 'pre-/pro-, mud',[4] and *morð* 'murder' which form a non-existing word (*for.morð*), instead of the correct parts *form* 'form' and *orð* 'word' (*form.orð*), which were, nota bene, also included in the list. All such irregularities were removed from the list, leaving 739 morphological parts, which were used as the basis for identifying terms from the text, retrieving words containing at least one morphological part (for example *form* or *orð* or even both). This gave the results shown in Table 2.

TABLE 2. A method based on a BiLSTM Compound Splitter for Icelandic

True positives:	93
False positives:	150
False negatives:	6
Precision:	38.3%
Recall:	93.9%
F_1:	0.544

As the numbers show, this resulted in a higher F_1 score, with a much higher recall but less precision, i.e., the recall is less accurate.

Comparing the two methods, we can see that the method that uses the compound splitter results in a much higher recall, meaning that a database containing known stems and morphological parts of a given field's vocabulary can be of great help when extracting new ones. So far, this has only been investigated with a single book chapter as test data, but

3 The majority of words in the Database of Icelandic Morphology consist of compounds: "Out of 278,764 paradigms [...] on Dec. 15th 2015, 32,118 entries were non-compounds, and the remaining 246,646 entries were compounds" (Bjarnadóttir 2017:14).

4 In Icelandic *for* can be used as a prefix 'pro, pre, etc.' or a noun meaning 'mud'.

the results are promising. It should be noted, however, that its precision is quite low, 38.3%, meaning it returns multiple false positives, which is disadvantageous as it can be time-consuming for editors to filter out the false positives.

The tf-idf method has neither high recall nor high precision, which stems from the fact that it returns only 111 candidates, compared to the 243 candidates the compound method suggests, and it suggests multiple words that cannot be considered true positives.

Finally, we combined the two methods, which resulted in a list containing only terms deemed probable by both, yielding the results shown in Table 3.

TABLE 3. The two methods referred to above combined

True positives:	46
False positives:	15
False negatives:	53
Precision:	75.4%
Recall:	46.5%
F_1:	0.575

The combined method yields interesting results: The recall is the same as for the tf-idf method, at 46.5%, but its precision, 75.4%, is the highest one, meaning that for every four words the method returns, three of them are correct.

5. Conclusions and future work

We need to look beyond the traditional (and perhaps somewhat dated methods) of manual termbase construction and try to simplify the process of preparing, storing, and sharing term glossaries. The automatic term extraction tool, built for the workbench, shows promising results. As noted, it is the first tool of its kind to support Icelandic, and terminology databases have until now been constructed by hand. As a result, our focus was on maximizing the tool's ability to gather potential new terminology and create a sizable initial database suitable for further computerized work and research. Accordingly, term recall was of primary importance and was heavily emphasized over precision during the tool's

development. Fine-tuning precision will be part of future work on TermPortal.

Furthermore, the numerous available terminology databases for Icelandic should be looked into in more detail and used as a basis for further development of the compound-splitter method. Moreover, once fully functional and voluminous enough, TermPortal's data can be used in various, useful ways, such as automatic indexing of scholarly and scientific work or automatic keyword extraction for all sorts of texts.

References

Barkarson, Starkaður, Steinþór Steingrímsson, Hildur Hafsteinsdóttir, Þórdís Dröfn Andrésdóttir, Inga Guðrún Eiríksdóttir & Bolli Magnússon 2021. *IGC-Journals-21.12* (*The Icelandic Gigaword Corpus – scholarly and scientific journals*), CLARIN-IS. <http://hdl.handle.net/20.500.12537/166>. Accessed August 2022.

Bjarnadóttir, Kristín 2017. Phrasal compounds in Modern Icelandic with reference to Icelandic word formation in general. In: Carola Trips & Jaklin Kornfilt (eds.). *Further investigations into the nature of phrasal compounding*. Berlin: Language Science Press, 13–48.

Daðason, Jón, David Erik Mollberg, Hrafn Loftsson & Kristín Bjarnadóttir 2020. Kvistur 2.0: a BiLSTM Compound Splitter for Icelandic. In: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.). *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 3991–3995.

Daníelsson, Hjalti, Ágústa Þorbergsdóttir, Steinþór Steingrímsson & Gunnar Thor Örnólfsson 2020. TermPortal: A Workbench for Automatic Term Extraction from Icelandic Texts. In: Béatrice Daille, Kyo Kageura & Ayla Rigouts Terryn (eds.). *Proceedings of the 6th International Workshop on Computational Terminology* (COMPUTERM 2020). Marseille: European Language Resources Association, 8–16.

Frantzi, Katerina T., Sophia Ananiadou & Hideki Mima 2000. Automatic recognition of multi-word terms: the *C-value/NC-value* method. *International Journal on Digital Libraries* 3, 115–130.

Ingólfsdóttir, Svanhvít Lilja, Hrafn Loftsson, Jón Friðrik Daðason & Kristín Bjarnadóttir 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In: Mareike Hartmann & Barbara Plank (eds.). *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku: Linköping University Electronic Press, 310–315.

Levenshtein, Vladimir Iosifovich 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10(8), 707–710.

Rögnvaldsson, Eiríkur 2013. *Hljóðkerfi og orðhlutakerfi íslensku*. Reykjavík. <https://notendur.hi.is/eirikur/hoi.pdf>. Accessed September 2022.

Steingrímsson, Steinþór, Örvar Kárason & Hrafn Loftsson 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In: Ruslan Mitkov & Galia Angelova (eds.). *Proceedings of Recent Advances in Natural Language Processing,* RANLP 2019. Varna: INCOMA Ltd., 1161–1168.