Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pages 8–16
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

TermPortal: A Workbench for Automatic Term Extraction from Icelandic Texts

Hjalti Daníelsson[1,2], Ágústa Þorbergsdóttir[2] Steinþór Steingrímsson[2], Gunnar Thor Örnólfsson[2]
[1]University of Iceland
[2]The Árni Magnússon Institute for Icelandic Studies
hjaltid@hi.is, agusta.thorbergsdottir@arnastofnun.is,
steinthor.steingrimsson@arnastofnun.is, gunnar.thor.ornolfsson@arnastofnun.is

Abstract

Automatic term extraction (ATE) from texts is critical for effective terminology work in small speech communities. We present TermPortal, a workbench for terminology work in Iceland, featuring the first ATE system for Icelandic. The tool facilitates standardization in terminology work in Iceland, as it exports data in standard formats in order to streamline gathering and distribution of the material. In the project we focus on the domain of finance in order to do be able to fulfill the needs of an important and large field. We present a comprehensive survey amongst the most prominent organizations in that field, the results of which emphasize the need for a good, up-to-date and accessible termbank and the willingness to use terms in Icelandic. Furthermore we present the ATE tool for Icelandic, which uses a variety of methods and shows great potential with a recall rate of up to 95% and a high C-value, indicating that it competently finds term candidates that are important to the input text.

Keywords: terminology extraction, corpora, Icelandic

1. Introduction

Terminology extraction is the task of automatically extracting relevant terms from a given corpus. An up-to-date reliable termbase of systematically collected terms or term candidates from recent texts is of great importance to translators and users of translated texts. Such a tool can be very useful for standardizing vocabulary in specialized fields, which again is crucial for translation work, leading to increased translator productivity and helping to make new texts and translations more coherent and unambiguous.

Until now, terminology databases in Iceland have been constructed manually by experts in their subject fields. Termbases in more than 60 different fields have been created and made available online in Íðorðabankinn[1]. The Translation Centre of the Ministry for Foreign Affairs has also made their terminology database available online[2].

There are several downsides to manual collection of terminology. New terminology often takes a long time to reach publicly accessible termbases. Some of the collected terms may not see widespread use before being supplanted by newer or better-known ones, but nonetheless linger on in the termbase, increasing the risk of ambiguity unbeknownst to the termbase editors. In certain fields there may also be a lack of experts interested in doing the terminology work, making standardization of terminology even harder. Through the adoption of state-of-the-art methods for the automatic extraction of terms, new terminology can be made available much earlier in publicly accessible termbases, where it can facilitate more effective standardization. Editors can also more easily collect statistics about terminology use and cite real-world usage examples.

We present TermPortal, the first build of a workbench for semi-automatic collection of Icelandic terminologies. The workbench includes an automated terminology extraction tool that provides editors of terminologies with lists of new terminology candidates from relevant texts. For our initial version we focus on the acquisition of potential new terms, and the domain of finance. An emphasis on recall over precision allows us to potentially create a corpus with which to conduct future research. Meanwhile, since financial terminology is used in a variety of different texts, there is abundant data on which to try our system – a useful property both for developing our system and for learning about how difficult it is to automatically extract terminology from different texts.

There is also a great need for a continually updated termbase in this active field, as was confirmed in a thorough survey conducted at the start of the project. We describe the methodology for the survey and the results in Section 3, while the emphasis on term acquisition over term filtering is noted in Section 4.

TermPortal consists of two main pieces of software: One is the TermPortal workbench described in Section 4, which includes an automatic pipeline to extract terminology from media and a web platform where users can create, manage and maintain termbases. The other is the Automatic Term Extraction (ATE) system described in Section 5. It is a central component in the TermPortal workbench, but can also be used in isolation.

2. Related Work

TermPortal is not the first termbase management tool to offer ATE, although it is the first to support Icelandic.

2.1. ATE Management

Tilde Terminology[3] is a cloud-based terminology extraction and management tool based on the Terminology as a Service (TaaS) project (Gornostay and Vasiljevs, 2014). It allows users to upload monolingual documents and employs the CollTerm term extraction system (Pinnis et al., 2012) to extract term candidates, as well as offering various methods for automatic identification of translations for the candidates, such as lookup in EuroTermBank (Rirdance, 2006; Gornostaja et al., 2018) and parallel data that

[1] http://idord.arnastofnun.is
[2] http://hugtakasafn.utn.stjr.is
[3] term.tilde.com

Tilde have mined from the web. There are also several multilingual terminology workbenches available on the web. Terminologue[4] is an open-source cloud-based terminology management tool developed by Dublin City University. It is quite flexible and enables users to define the languages used in a given termbase, as well as employing a hierarchical structure for terminology domains. It also supports importing and exporting termbases in TBX format. A multitude of commercial solutions is also available. Among the solutions available are SDL MultiTerm, TermWiki and Termbases.eu.

In our work, we sacrifice some of the flexibility provided by workbenches such as Terminologue for the sake of making the process of extracting the terms themselves as straightforward and linear as possible. Much like in Tilde Terminology, we offer ATE as well as lookup in an existing termbank[5], but do not support term alignment between languages in the current version.

2.2. Automatic Extraction

While there are no studies on automatic extraction specifically for Icelandic, much less a particular domain such as finance, terminology extraction from monolingual corpora is a well-established field applying many different approaches. It can be said that there are two general approaches to automated terminology extraction from monolingual texts: statistical and rule-based. The rule-based methods commonly include tokenization, PoS-tagging, lemmatization, stemming and other common Natural Language Programming (NLP) approaches to linguistic analysis. A number of tools support these approaches for Icelandic texts: Some are multifunctional, such as Reynir (Þorsteinsson et al., 2019), and the IceNLP collection (Loftsson and Rögnvaldsson, 2007), while others are specialized in particular tasks: ABLTagger, a new BiLSTM PoS-tagger, has reached 95.17% accuracy for PoS-tagging Icelandic texts with a rich morphosyntactic tagset (Steingrímsson et al., 2019); and a recent lemmatization tool, Nefnir, has shown good results for lemmatizing Icelandic texts (Ingólfsdóttir et al., 2019). Some of these tools are employed in our extraction process.

In terminology extraction, linguistic features specific to an individual language are commonly used, in particular morphosyntactic information and discourse properties that distinguish noun phrases which are technical terms (Justeson and Katz, 1995). The statistical methods range from very basic approaches like counting the number of n-grams in a document to using statistical measures and recently word embeddings in neural networks. The use of statistical measures in automated term extraction can be governed by two aspects of terms, termhood and unithood, introduced by Kageura and Umino (1996). Termhood is considered to be "the degree to which a stable lexical unit is related to some domain-specific concepts", a notion closely related to a standard definition of a term. Meanwhile, unithood is "the degree of strength or stability of syntagmatic combinations and collocations". Certain methods can exploit both unithood and termhood, the C-value introduced by Frantzi et al. (2000) being a common measure. Many successful systems, however, employ a hybrid approach, using linguistic features to limit the search space and then applying statistical filtering. See Vintar (2010) for an example of such an approach. More recently deep learning approaches have been tested, using word embeddings. Zhang et al. (2018) give an example of such a method, using word embeddings to compute 'semantic importance' scores for candidate terms, which are then used to revise the scores of candidate terms computed by a base algorithm using linguistic rules and statistical heuristics. Bilingual extraction is another approach to ATE. In contrast to monolingual extraction which is concerned with identifying terms in a corpus, bilingual extraction primarily deals with aligning terms in different languages. Recently some advances have been made in automatically extracting terms from comparable corpora using deep learning methods (Liu et al., 2018; Heyman et al., 2018). The general idea in these methods is to project word embeddings in both languages to a shared vector space. In our work, however, we focus on monolingual extraction.

3. The Survey

Given the apparent scarcity of up-to-date terminology databases in Iceland, the first part of our project was to examine the views domain specialists hold on terminology. More specifically, we wanted to investigate the perceived value of terminological data, the level of interest in the use, acquisition, and sharing of terminology, the quality of facilities currently employed for storage of term databases, and the level of importance assigned to instant access to current terminological data.

In order to be better able to deliver a useful system, we decided to work on only one domain for the first version of our system, the domain of finance. There were several reasons for the choice of this particular domain. While term collections in any domain require regular updates to prevent their obsolescence, Iceland's financial environment has changed extensively in recent times. The introduction of European directives alone has brought a host of new concepts to the field. Extant Icelandic terminology databases and dictionaries now contain a great number of deprecated, obsolete and superseded terms, making it even more difficult to find the correct Icelandic financial terms in what is already a relatively complex field for terminology.

By narrowing our focus we are also able to get a comprehensive view of the term usage and needs of a specific group. We therefore commissioned a survey on the subject of terminology within this domain, and submitted it to financial institutions, corporations, and translation agencies.

3.1. How the Survey was Formulated

The survey included questions on term-related issues, term cataloging, and opinions on terminology and term-related tasks, including the importance of terms in the workplace and the willingness to share collected terminology.

[4] www.terminologue.org

[5] The Icelandic Term Bank: https://idord.is, https://clarin.is/en/resources/termbank/

Question	Very High	High	Neutral	Low	Very Low	None
Importance of term translations	60.9%	21.7%	17.4%	0.0%	0.0%	0.0%
Interest in free access to a trustworthy termbank	69.7%	26.1%	0.0%	4.2%	0.0%	0.0%
Willingness to share terminology with others through a termbank	50.0%	22.7%	22.7%	0.0%	0.0%	4.5%
Willingness to take part in terminology work with others	27.3%	27.3%	36.3%	9.1%	0.0%	0.0%

Table 1: Questions in survey about interest in using and working towards a common termbase.

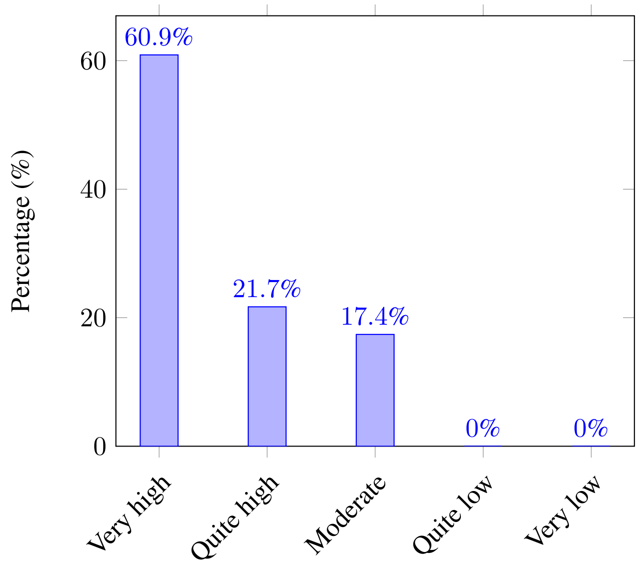


Figure 1: Are term translations of high or low importance?

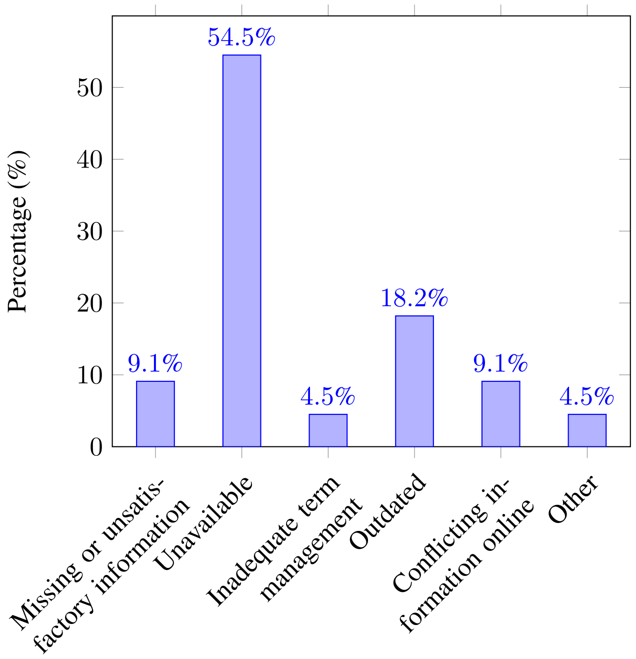


Figure 2: Most pressing issues related to terminology in the workplace.

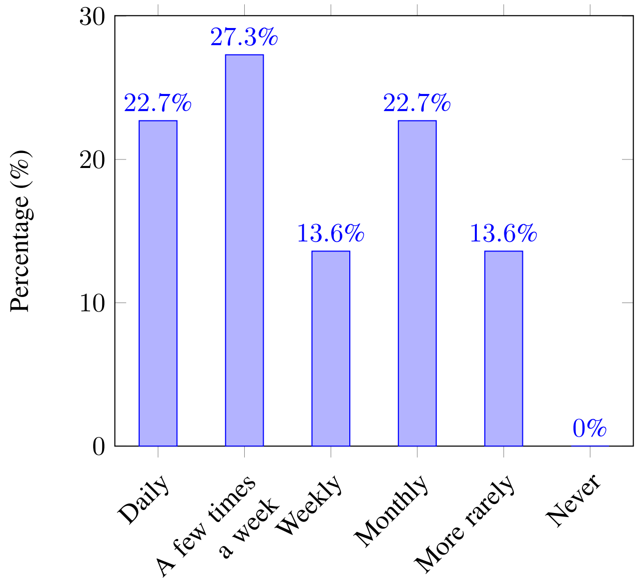


Figure 3: How often do participants look up domain terms?

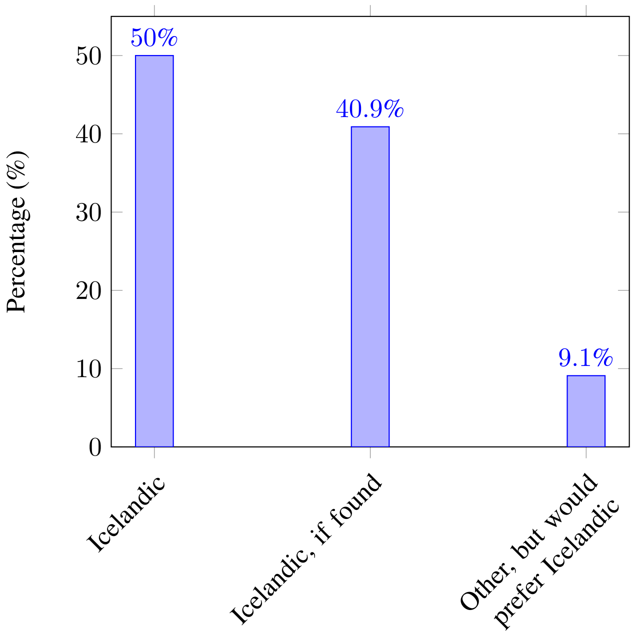


Figure 4: What is the language of choice for terminology?

3.2. Survey Results

Since the subject matter was so clearly delineated – terminology within a single domain – the survey was only directed at the most prominent organizations in that particular field. For each participating organization, the survey was put to the single representative considered to have the most extensive experience and play the most significant role in that organization's approach to, and policies on, terminology. The number of invited participants was consequently kept fairly low, but was also estimated to represent those organizations that would have the greatest interest in the potential value of terminology within the field, and the most extensive abilities to deploy that terminology in everyday tasks. As a result, their opinions on the subject were considered highly relevant and extremely valuable. Out of the twenty-five invited organizations, twenty-three took part in the survey – a response rate which the survey conductors considered to be quite high and thus likely to result in more reliable survey results. Moreover, the conductors noted that the representatives' extensive experience within their respective organizations was likely to produce informed, thought-out answers that could be trusted to be truthful; even more so since the survey was anonymous and conducted through an intermediary rather than The Árni Magnússon Institute for Icelandic Studies directly. The only overt classification of participants was a grouping of answers into the three categories mentioned earlier: institutions, corporations and translation agencies.

Participants were asked sixteen questions. The results were decisive, and markedly in the terms' favor. Table 1 shows the responses to four of the questions, those concerning interest in using a common termbase and willingness to take part in building one. The survey participants see definite value in domain terminology with almost everyone in favor of free access to a trustworthy termbank and the majority interested or willing to take part in terminology work. Table 1 also displays a notable downward gradient among the percentage of responses in the Very High column: There is clear interest among participants in having access to high-quality terminology, but slightly less so in participating in information sharing with others (including potential competitors), and rather less so in devoting time and manpower of their own to create a terminology collection at all. Of the three types of participating organizations, translation agencies - which tend to have the smallest staff - were the ones with the lowest willingness to share their own data and take on additional work load. This puts The Árni Magnússon Institute for Icelandic Studies at an advantage, being an institute whose domain is separate from the survey participants and whose staff includes experts knowledgeable in this field: It indicates that if we were to lay the terminology groundwork by establishing TermPortal, we would have gotten past any major hurdles of cooperation from these participants, and could likely expect a higher willingness in active participation (such as through user testing) during future stages of the tool's development.

Access to terminology and term databases was deemed both of clear importance (see Figure 1) and, in its current form, severely lacking (see Figure 2). Also, even though the majority of respondents estimated that their staff look up domain terms weekly or more often (see Figure 3), most participants responded that no term registration whatsoever was performed within their organization. At the same time, a majority believed the most pressing issue related to terminology in the workplace was that up-to-date terms had not been collected and made available to all (see Figure 2).

This lack of availability of Icelandic-language collections, both for up-to-date terms and in general, was reinforced when the organizations were asked where their staff would look for assistance with translations of finance terminology. A majority responded that they would ask their coworkers, rather than look to online resources, specialists in the field or any other potential resource.

Attitudes toward Icelandic terminology in particular were predominantly positive. As evident from Figure 4, when asked about their chosen language for terminology, none of the participants explicitly said they preferred non-Icelandic terms and over 90% stated they used Icelandic terminology, either when available or exclusively.

These results clearly indicate the importance of easy and open access to up-to-date data. Indeed, the availability of Icelandic terminology may be seen as a vital precondition for clear and efficient communication in each field.

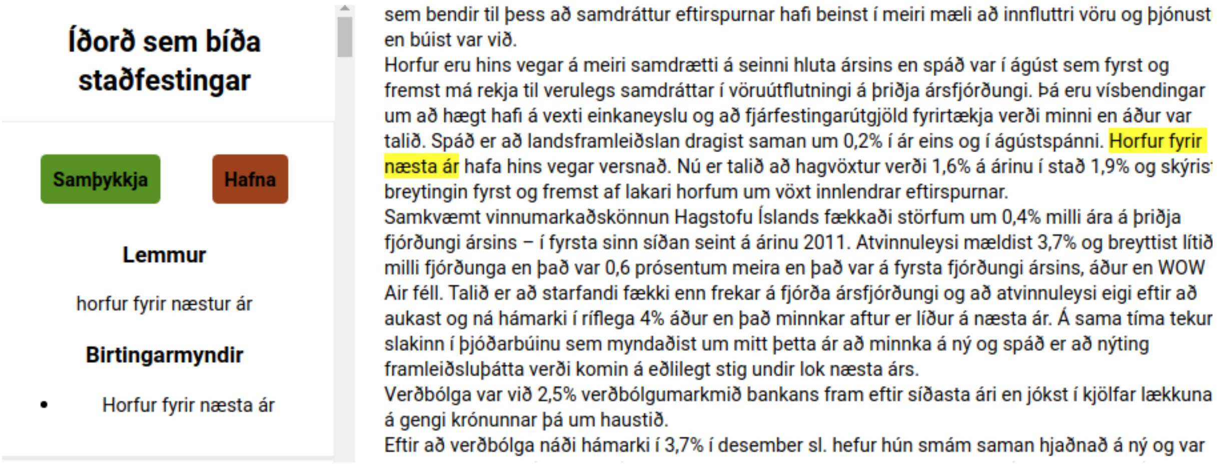
To be able to meet the needs of this influential user group, and other professional terminology users, exploring new ways of implementing terminology collection and storage was necessary. We need to look beyond the increasingly dated methods of manual termbase construction and try to simplify the process of preparing, storing, and sharing term glossaries. This will enable specialists in the field to focus on more productive endeavors than basic terminology work and make it easier to centralize that work. Our aim was that all potential users of Icelandic terminology, including field specialists and translators, would be able to spend less time hunting down possible term candidates, and instead could simply edit or approve listed candidates, ensuring greater consistency in terminology use, dissemination, standardization, and translation.

4. The TermPortal Workbench

The TermPortal workbench is an online terminology acquisition and management system. Authenticated users can create termbases and upload texts which are subsequently processed by the automatic term extraction (ATE) tool described in Section 5. After candidate terms are extracted, they are displayed alongside the source text. Selecting a term candidate highlights each of its occurrences in the text, allowing the user to quickly see the phrase in context. An example of this is shown in Figure 5. Furthermore, each occurrence's enveloping sentence is stored, for later use as usage examples.

Term candidates have five defined stages:

- Automatically extracted
- Rejected
- Manually entered / Accepted
- Reviewed
- Publishable



sem bendir til þess að samdráttur eftirspurnar hafi beinst í meiri mæli að innfluttri vöru og þjónust
en búist var við.

Horfur eru hins vegar á meiri samdrætti á seinni hluta ársins en spáð var í ágúst sem fyrst og fremst má rekja til verulegs samdráttar í vöruútflutningi á þriðja ársfjórðungi. Þá eru vísbendingar um að hægt hafi á vexti einkaneyslu og að fjárfestingarútgjöld fyrirtækja verði minni en áður var talið. Spáð er að landsframleiðslan dragist saman um 0,2% í ár eins og í ágústspánni. Horfur fyrir næsta ár hafa hins vegar versnað. Nú er talið að hagvöxtur verði 1,6% á árinu í stað 1,9% og skýris
breytingin fyrst og fremst af lakari horfum um vöxt innlendrar eftirspurnar.

Samkvæmt vinnumarkaðskönnun Hagstofu Íslands fækkaði störfum um 0,4% milli ára á þriðja fjórðungi ársins – í fyrsta sinn síðan seint á árinu 2011. Atvinnuleysi mældist 3,7% og breyttist lítið milli fjórðunga en það var 0,6 prósentum meira en það var á fyrsta fjórðungi ársins, áður en WOW Air féll. Talið er að starfandi fækki enn frekar á fjórða ársfjórðungi og að atvinnuleysi eigi eftir að aukast og ná hámarki í ríflega 4% áður en það minnkar aftur er líður á næsta ár. Á sama tíma tekur slakinn í þjóðarbúinu sem myndaðist um mitt þetta ár að minnka á ný og spáð er að nýting framleiðsluþátta verði komin á eðlilegt stig undir lok næsta árs.

Verðbólga var við 2,5% verðbólgumarkmið bankans fram eftir síðasta ári en jókst í kjölfar lækkuna
á gengi krónunnar þá um haustið.

Eftir að verðbólga náði hámarki í 3,7% í desember sl. hefur hún smám saman hjaðnað á ný og var

Figure 5: A term candidate highlighted in context in the TermPortal workbench.

At this stage the user can either accept or reject the term candidate. Rejected term candidates will be hidden from future ATE results for that termbase, but can be viewed separately and recovered.

The tool extracts term candidates and makes note of each candidate's occurrences in the source text, highlighting them on command. Users can then choose whether to accept or reject each of the term candidates provided by said tool. Accepted candidates are added to the active termbase, and can be further processed, adding definitions, references to related terms and translations. Lists of fully or partially processed terms can be exported in standard formats such as TBX[6] and CSV[7], enabling easy integration of those termbases into other systems which conform to those standards. In addition to supporting exportation of termbases, users can share them with other users with varying privileges.

- *Owners*, or co-owners have full privileges over the termbase, including giving other users privileges and general termbase administration in addition to managing the terms within the termbase. The user who creates a termbase is by default its owner.
- *Editors* have privileges over terms and texts within the termbase, but not the termbase itself. Their privileges include uploading and processing texts, accepting or rejecting term candidates, and modifying term entries.
- *Reviewers* can supply commentary on terms which have previously been accepted by *editors* or *owners*. They also have rights to mark terms as 'reviewed'.
- *Viewers* have view-only rights to the termbase and no edit privileges.

Until now, no publicly available workbench designed for terminology work in Icelandic has existed, meaning that editors for each domain set their own individual workflows and standards, which can cause difficulty when termbases are combined and centralized. A standardized work environment for terminology collection will enforce homogeneity in termbase structure between subject areas, facilitating easy termbase compilation. Interactive use of the ATE component turns the complex task of identifying new terms into a sequence of binary questions, greatly simplifying the workflow of termbase editors and potentially increasing productivity.

5. Automatic Term Extraction

The ATE tool lies at the heart of the TermPortal workbench. It accepts input in the form of Icelandic text, processes the text in order to find possible candidates, calculates the candidates' term likelihoods, and outputs a sorted list of those terms it deems most likely to be heretofore unseen terms within a given domain.

As noted, this 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 considered to be 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.

5.1. Data Preparation

Although the ultimate goal of the ATE tool is to be capable of handling texts from any domain, we initially focus on the financial sector as we do in other parts of the project. This means that we sourced testing data solely from that particular field. The data came in two forms: Randomly selected texts originating from various sources in the financial sector, primarily laws, regulations, reports, and educational materials; and known finance terms listed in the aforementioned terminology database compiled by the Translation Centre of the Ministry for Foreign Affairs (see Section 1). While the random compilation of the general texts – some of which carry confidentiality clauses – makes it impractical to publish them as datasets, all the known finance terms may be accessed through the Ministry's website, which allows content filtering according to subject area.

[6] ISO 30042:2019

[7] RFC 4180

Test Set	Random Clauses	Known Terms	Total
1	250	250	500
2	500	500	1,000
3	1,000	1,000	2,000
4	2,000	2,000	4,000

Table 2: The four test sets.

In order to evaluate the tool, we created four text files that combined these two types of data. Each file contained one sentence or clause per line and had an equal ratio between lines of random clauses and lines of known terms. The smallest file of 500 lines thus contained 250 random clauses and an additional 250 known terms; and with each test set the file size doubled, as shown in Table 2. During each test, one of these files served as the program's main input. Alongside that file, we provided the tool with two others: A unique list of just over 2,000 known finance terms, in lemmatized form, whose contents did not overlap with the terms added to the input file, and a list of 280 grammatical category patterns that corresponded to all known financial terms. Each entry in the pattern list contained an ordered sequence of grammatical tags, such as ('a', 'v', 'n'), corresponding to ('adjective', 'verb', 'noun'). In section 5.3 we describe how these patterns are used to identify potential term candidates in the program's input.

5.2. Methods for term extraction

In choosing our methods, we needed to consider certain constraints while trying to provide maximum coverage. The Icelandic language is morphosyntactically rich, with a relatively free sentence word order, high inflectional complexity, and a high ratio of compound words, all of which affect the linguistic aspects of term extraction (Bjarnadóttir et al., 2019). Moreover, while we focused on the financial sector during development, the ATE tool needed to be domain-agnostic by design and be able to run without any prior training, which already eliminates a host of options. Lastly, certain supplementary data, in the form of known terms and stop-words from that domain, might be available at times but could not be a prerequisite. As a result of these factors, we implemented three methods of term extraction, all of which are applied to input that the tool has already lemmatized.

The first method is C-value (Frantzi et al., 2000), modified to include single-word terms (Barrón-Cedeno et al., 2009). This is likely one of the best-known term search methods in existence. It is language- and domain-independent, does not require any information other than the text input itself, and relies on the kind of linguistic preprocessing (i.e. tagging and category filtering) that would likely always be incorporated by any ATE tool when applied to Icelandic texts.

The second method, which we term the 'stem ratio', is one we created specifically for this particular project, and is intended to take advantage of the high number of compound words in Icelandic while remaining unaffected by the issue of multiple potential word orders. When applying this method, the ATE tool employs a separate program called Kvistur, which decompounds Icelandic words (Daðason and Bjarnadóttir, 2015). Icelandic compounds are morpohologically right-headed (Bjarnadóttir, 2017), so through Kvistur the ATE tool analyzes the morphological structure of the words contained within each term candidate, extracts all rightmost stems, and compares them to all rightmost stems found in known candidates. If the total number of all the stems in the words of a given candidate is A, and the total number of those same stems in the entirety of known candidates is B, the stem ratio for that candidate is A/B. (Candidates with no compound words are simply not assigned stem ratios.) Hence, the more common that a candidate's morphological heads are within known terms, the higher its stem ratio will be. A candidate with a high stem ratio shares a great deal of both morphological structure and meaning with existing terms, and is itself thus likely to be a new term. It should be noted that Icelandic is a fairly complex language; as such, we will constrain our discussion of ATE methods to ways in which this particular project was implemented, since any further details would require a separate chapter unto themselves.

The third method is Levenshtein-distance, which in our context is the minimum number of single-character edits required to change one string into another. The Levenshtein algorithm is comparatively straightforward, well supported in Python, and has been used or considered for ATE (Nazarenko and Zargayouna, 2009; Droppo and Acero, 2010) and other term-extraction-type projects in the past (Runkler and Bezdek, 2000). For each candidate, we find the lowest possible Levenshtein-distance between it and any known term. The lower this value is, the more the candidate resembles a known term letter-for-letter, irrespective of factors such as multiple inflections or morphosyntactic structures.

Overall, these three methods cover a wide range of possible terms. The C-value finds those candidates that are clearly important to the input itself, in terms of their unithood and termhood. The stem ratio finds any candidates – generally lengthy and complex ones – whose composition, structure and meaningful parts bear clear resemblance to existing terms, even when the less-meaningful parts may be completely dissimilar. Lastly, Levenshtein-distance takes a much rawer approach and finds those candidates – here generally ones that are short or contain simple words – which simply resemble known ones in terms of spelling, and which might otherwise be overlooked by the stem ratio. It should be noted that since the latter two methods rely on comparisons to the list of known terms, they will not be adversely affected by candidates' low frequencies of occurrence in the input – but they do require an actual list of terms in order to work at all. In addition, variations on all three approaches are certainly possible, but for the most part we decided to refrain from complicating our algorithms until we'd compiled a solid database of terms for further testing; the one exception being a slight change to C-value calculations to account for single-word terms (Barrón-Cedeno et al., 2009).

Linguistic Processing Tool	Set size	C-value	L-distance	S-ratio	Recall (%)
ABLTagger + Nefnir		1.744	7.080	22.746	80.00
Reynir	500	1.820	7.676	23.554	92.80
ABLTagger + Nefnir		1.735	6.903	22.930	83.20
Reynir	1,000	1.773	7.247	23.031	92.40
ABLTagger + Nefnir		2.115	6.883	20.254	84.90
Reynir	2,000	2.238	7.315	20.342	89.60
ABLTagger + Nefnir		2.433	7.101	19.538	89.35
Reynir	4,000	2.501	7.421	20.041	89.00

Table 3: Average values for ATE methods across linguistic processors for all data sets. Threshold values not applied.

5.3. Usage

The tool is divided into the following four sections, which run sequentially: Preprocessing, linguistic processing, statistical processing, and output.

During preprocessing, the tool checks what data is being supplied – the main input and a list of category patterns are mandatory for every activation, while a list of known terms and a separate list of stop words (not used in our tests) are optional – and loads any comparison data into memory. If a list of known terms is included, its contents are expected to be in lemmatized form for comparison purposes, although if need be the ATE tool itself is capable of lemmatizing the list's contents by using the linguistic support programs detailed below. If the tool is provided with such a list it will also, through the aforementioned program Kvistur, compile an additional list containing the compound word heads of all known terms.

In the linguistic section, the tool reviews one line at a time, and tokenizes, tags and lemmatizes it. There are two primary ways in which the tagging and lemmatization may be done, decided on by the tool's administrator: Through the Reynir[8] Python package (Þorsteinsson et al., 2019), or through the tagger ABLTagger[9] (Steingrímsson et al., 2019) and lemmatizer Nefnir[10] (Ingólfsdóttir et al., 2019). For a language as morphologically rich as Icelandic, we felt it necessary to have more than one processing option, although it should be stated that our purpose is not to compare the programs themselves – the primary notable difference is that Reynir automatically performs a more exhaustive and thus more time-consuming analysis. At the end of this section, each line has been converted to a sequence of tuples, where each tuple contains a single, now lemmatized word from the phrase, and a corresponding tag for that word's grammatical category.

One last function serves as a bridge to the statistical section: Before the ATE tool applies any of the three extraction methods, it compares each tuple sequence against every single entry in the list of grammatical category patterns known to represent known terms. Any continuous part of the sequence that matches a known grammatical word pattern is automatically added to the list of term candidates. The extraction methods – C-value, stem ratio and Levenshtein-distance – are then calculated for every entry on that candidate list.

In the output section, the tool prepares the list for use by subsequent parts of TermPortal. The tool also includes threshold values that may be set for each of the three methods, in which case every candidate will have to meet at least one of the thresholds (if applicable, since not all methods are necessarily being applied each time) in order to remain on the candidate list at all. ATE programs generally require specialist input when the final term lists are reviewed. As such, these threshold values help keep the output manageable, particularly while the tool's focus is still on recall.

5.4. Evaluation

In a project of this nature – where the ATE tool will be applied to an input of continuously changing size and content, rather than a predetermined corpus – the primary focus of evaluation is whether the tool demonstrably works against test inputs of, again, varied sizes and content: If it can properly parse the input, compare it against known terms, find the majority of candidates we know to be present, and display sensible statistical values over a spectrum of different inputs. This effectively means we wanted to measure its recall of the candidates we had intentionally inserted. As noted earlier, measuring precision, on the other hand, was not considered a priority at this stage of development. For a similar reason, we do not focus on narrowing the threshold values during this initial run; rather, we expect to continually adjust them once the TermPortal is actively receiving live data and compiling terminology. Instead, we want to see if the values are being applied in a consistent manner during these initial runs.

The results from our four datasets were consistent and promising, as may be seen in Table 3. Therein, we see the results of applying the two linguistic processing methods and subsequently the three statistical processing methods (C-value, Levenshtein-distance and Stem ratio) to the four data sets described in Table 2 (in each case, the terms being measured for recall were removed from the list of known terms that the program used to calculate Levenshtein-distance and Stem ratio). The lowest recall percentage, 80.0%, resulted from the smallest dataset when parsed by ABLTagger and Nefnir, while Reynir had 92.8% on that same dataset. Larger datasets increased overall recall for ABLTagger and Nefnir with both models, reaching 89.35% on the largest set, while Reynir's lowest recall dipped only to 89.0% with that same set. As may be seen, once the amount of input reaches a particular threshold, the

[8] https://pypi.org/project/reynir

[9] https://github.com/steinst/ABLTagger

[10] https://github.com/jonfd/Nefnir

recall rates between the two processing options tend to converge.

Averages for the values calculated by the statistical methods – C-value, Levenshtein-distance and stem ratio – were assigned to every possible candidate, not merely the ones on our recall list, and were highly consistent across both linguistic options for every dataset. The highest difference in averages was 0.123 for C-value in the 2,000-line set, 0.596 for Levenshtein-distance in the 500-line set, and 0.808 for stem ratio in the 500-line set.

Lastly, it should be noted that between them, these linguistic processing tools collectively managed impressive recall. In fact, out of the 2,000 known terms we inserted into the largest dataset, only 100 failed to be acknowledged at all. Given that many of the financial terms contain complex words that may at times be quite dissimilar from most text that the linguistic programs were trained on or programmed to recognize, a collective 95% recall rate – meaning that in at least one of the two processing options, the words were correctly tokenized, tagged, lemmatized, matched to known grammatical category patterns, and passed on for value calculation – is a highly positive result. As noted earlier, the two options offer differing depths of linguistic processing, with the associated increase in workload and processing time. As such, Reynir is likely to be used more often on shorter texts, particularly if a more thorough approach is required, while ABLTagger and Nefnir are the preferred choice for processing greater volumes of incoming text at a reasonable pace.

6. Availability and licensing

The TermPortal is in closed testing. It will be open for use for all parties interested in undertaking terminological work in Iceland, running on servers at The Árni Magnússon Institute for Icelandic Studies[11]. The ATE tool is available under an open Apache 2.0 license.

7. Conclusion and Future Work

We have presented TermPortal, a workbench for terminology work using an automated term extraction tool, adapted to Icelandic and the domain of finance. The automatic term extraction tool, built for the workbench, shows promising results with a recall rate of up to 95%. The workbench and the ATE tool show great potential in answering the needs of industry, as manifested in a survey we conducted among the most prominent user group, which shows great interest in improving the state of affairs in Icelandic terminology work within the field of finance.

We have implemented approaches to term extraction suitable to data at hand. As more data accrues we expect to develop a far more robust test set than the one used for our initial tests. This will permit greater granularity of test results, along with variations such as testing other term ratios than 50/50 in the program's input. Other future work may include using deep learning approaches, such as word embeddings and bilingual extraction where parallel data is available. To improve the workbench, prospective users will be involved in testing and the resulting feedback used to help adapt the system even further to the needs of users. Users are also expected to help test the quality of our term databases, with an eye toward improving the precision with which the ATE tool collects new terms. Furthermore, user testing will yield precision statistics for the ATE tool, enabling us to tweak the parameters of the system to give a good balance of precision and recall.

8. Bibliographical References

Barrón-Cedeno, A., Sierra, G., Drouin, P., and Ananiadou, S. (2009). An improved automatic term recognition method for Spanish. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 125–136. Springer.

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

Bjarnadóttir, K. (2017). Phrasal compounds in modern Icelandic with reference to Icelandic word formation in general. In *Further investigations into the nature of phrasal compounding*, pages 13–48. Language Science Press, Berlin.

Daðason, J. F. and Bjarnadóttir, K. (2015). Kvistur: Vélræn stofnhlutagreining samsettra orða. *Orð og tunga*, 17:115–132.

Droppo, J. and Acero, A. (2010). Context dependent phonetic string edit distance for automatic speech recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4358–4361. IEEE.

Frantzi, K. T., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3:115–130.

Gornostaja, T., Auksoriūtė, A., Dahlberg, S., Domeij, R., van Dorrestein, M., Hallberg, K., Henriksen, L., Kallas, J., Krek, S., Lagzdiņš, A., et al. (2018). eTranslation TermBank: stimulating the collection of terminological resources for automated translation. In *Proceedings of the XVIII EURALEX International Congress*, EURALEX 2018, Ljubljana, Slovenia.

Gornostay, T. and Vasiljevs, A. (2014). Terminology resources and terminology work benefit from cloud services. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC 2014, Reykjavik, Iceland.

Heyman, G., Vulic, I., and Moens, M.-F. (2018). A deep learning approach to bilingual lexicon induction in the biomedical domain. In *BMC Bioinformatics*.

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*, NODALIDA 2019, Turku, Finland.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.

[11] https://termportal.arnastofnun.is

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: A review.

Liu, J., Morin, E., and Saldarriaga, S. P. (2018). Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING 2018, Santa Fe, New Mexico.

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

Nazarenko, A. and Zargayouna, H. (2009). Evaluating term extraction.

Þorsteinsson, V., Óladóttir, H., and Loftsson, H. (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)*, pages 1397–1404.

Pinnis, M., Ljubešic, N., Stefanescu, D., Skadina, I., Tadic, M., and Gornostay, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering*, TKE 2012, Madrid, Spain.

Rirdance, S. (2006). *Towards Consolidation of European Terminology Resources: Experience and Recommendations from EuroTermBank Project*. Tilde.

Runkler, T. A. and Bezdek, J. C. (2000). Automatic keyword extraction with relational clustering and Levenshtein distances. In *Ninth IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2000 (Cat. No. 00CH37063)*, volume 2, pages 636–640. IEEE.

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

Vintar, S. (2010). Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16:141–158.

Zhang, Z., Gao, J., and Ciravegna, F. (2018). SemRe-Rank: Improving automatic term extraction by incorporating semantic relatedness with personalised PageRank. *TKDD*, 12:57:1–57:41.