ISLEX – a Multilingual Web Dictionary

Thórdís Úlfarsdóttir
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
E-mail: disa@hi.is

Abstract

ISLEX is a multilingual Scandinavian dictionary, with Icelandic as a source language and Danish, Norwegian, Swedish, Faroese and Finnish as target languages. Within ISLEX are in fact contained several independent, bilingual dictionaries. While Faroese and Finnish are still under construction, the other languages were opened to the public on the web in November 2011. The use of the dictionary is free of charge and it has been extremely well received by its users. The result of the project is threefold. Firstly, some long awaited Icelandic-Scandinavian dictionaries have been published on the digital medium. Secondly, the project has been an important experience in Nordic language collaboration by jointly building such a work in six countries simultaneously, by academic institutions in Iceland, Denmark, Norway, Sweden, The Faroe Islands and Finland. Thirdly, the work has resulted in a compilation of structured linguistic data of the Nordic languages. This data is suitable for use in further lexicographic work and in various language technology projects.

Keywords: lexicography, multilingual dictionary, Icelandic

1. Introduction

ISLEX is an online multilingual dictionary between modern Icelandic and the Scandinavian target languages (TLs) Danish, Norwegian (two standards, Bokmål and Nynorsk), Swedish, Faroese and Finnish. These languages all belong to the North Germanic language group and are closely related, except Finnish which belongs to the Finno-Ugrik language family. For convenience, the terms 'Scandinavian' and 'Nordic' in this paper include Iceland, Denmark and the Faroe Islands as well as the mainland countries Norway, Sweden and Finland.

The source language (SL) of ISLEX is Icelandic. It is a language spoken by only 320,000 people, but considered to be of some historical importance as it is still very close to Old Norse, the common ancestor of the Nordic languages of the Germanic group. It has rich morphology, nouns are inflected in four cases and have up to 16 forms. Both modern Icelandic and Old Norse are widely studied and researched in universities in many countries.

There is a long history of a formal cooperation between the Scandinavian countries on many levels. Their various joint actions are comprehensive, not least on a political and cultural level[1]. These nations have much in common and their collaboration is based on common values and similar cultures, the result of which is that establishing new inter-Nordic projects is relatively easy. There are certain Nordic funds which aim to strengthen and develop the cultural relations between the countries, and special programmes for Nordic inter-understanding, language teaching and lexicography. Such a solid foundation was important for the ISLEX project to be realized, as dictionary work is notoriously labour consuming and therefore costly. ISLEX has for the most part been funded by the governments of the participating countries, as well as by important funds such as *Nordplus Sprog* and *Nordisk Kulturfond*.

1 Formal Nordic collaboration stretches back to before the founding of the European Economic Community in 1957, later replaced by the European Union.

2. Project Background

The ISLEX project was initiated in the year 2005. At that time there was a great need for modern, up-to-date dictionaries between Icelandic and the other Scandinavian languages. The existing dictionaries (printed books) were nearly all (too) old and outdated and most of them were not considered to be up to the standard of the day. This was a good time to begin such a project as a new generation of web dictionaries had emerged around the turn of the century. At the time, most or all E-dictionaries of the Nordic languages were "retro-digitized", i.e. earlier printed works which had been made digitally available. ISLEX was from the outset designed for the web and was as such an innovation in lexicography (in this part of the world at least).

Originally ISLEX was organised as a joint project of three academic institutes, in Iceland, Norway and Sweden, but soon Denmark joined the group. As the work continued, some key people in the Faroe Islands and Finland showed interest in becoming partners in the dictionary with their respective languages, Faroese and Finnish.

The project is organised in the following way. The largest editorial staff, responsible for the description of the Icelandic source language and for the development and maintenance of the database, is located at *The Árni Magnússon Institute for Icelandic Studies* at the

University of Iceland, Reykjavík. Its partners, responsible for the Scandinavian TLs are:

The Society for Danish Language and Literature (DSL) in Copenhagen (for Danish),
The University of Bergen, Norway (for Norwegian),
The University of Gothenburg, Sweden (for Swedish),
The University of the Faroe Islands (for Faroese),
Helsinki University (for Finnish).

The target groups of ISLEX are Icelandic speakers with medium to advanced knowledge of the other Scandinavian languages, and Scandinavians with limited to advanced knowledge of Icelandic, e.g. teachers, students and translators. In fact, all those who need a modern dictionary between Icelandic and the other Nordic languages. The dictionary is thus intended to be bifunctional, which is necessary in the case of Icelandic as such a small language society does not have the capacity for publishing many kinds of dictionaries, aimed at different target groups (cf. Sanders, 2005).

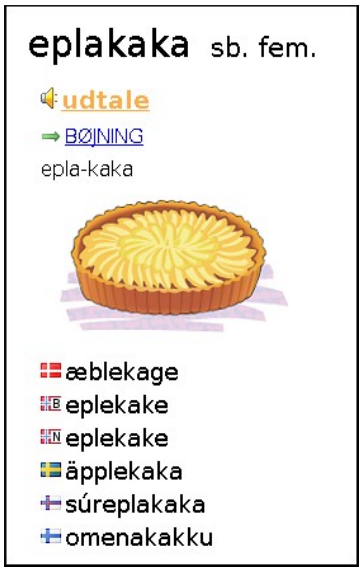

Figure 1: *Eplakaka* 'apple cake'. The simplest type of an entry in ISLEX with its six target languages.

ISLEX was opened to the public in November 2011, even if some parts of it were not completed at the time. The work on Icelandic, Danish, Norwegian and Swedish has since been (mostly) finished, but Faroese and Finnish are still in progress and have not yet (in March 2014) been opened on the web.

3. Working Method

At the beginning of the dictionary project it became evident that a new description of the Icelandic source language was absolutely necessary for the work ahead, even if ISLEX was going to be a bilingual dictionary (or several bilingual dictionaries). The detailed processing of the SL had to be done mainly because not the same people worked on the SL and the TLs, so every element in the SL had to be as clear-cut as possible. Besides, many words are polysemic (i.e. have more than one meaning, like the English words *bar* and *base*), and these had to be sorted out and divided into meanings. Cross references had to be made, examples had to be composed, and so on. In the process, a monolingual, corpus-based Icelandic dictionary was being compiled, almost accidentally, as a by-product of the preparatory work.[2]

Apart from this, in the early stages the vocabulary was divided into about 650 semantic fields. Articles belonging to a certain field were processed at the same time as a bundle, to ensure better consistency throughout the dictionary. This method has been used both by the editors of the SL and the TLs instead of working through the alphabet from A to Ö.

The inclusion of several target languages is not without its complications. The work has to be well coordinated between the institutions to avoid possible conflicts. And from the TLs' point of view, since the Scandinavian languages have so much in common in large proportions of the core vocabulary (except Finnish), it may be tempting to choose an equivalent that has the same etymological stem, but this may not be a true equivalent in many cases. The lexicographic term 'false friend' often comes to mind. However, the database and the editorial environment allow the TL editors to include detailed information on their chosen equivalents, in the process of translating the lexical units (Sigurðardóttir *et al.,* 2008). Each of the TLs have their own editing style, e.g. regarding the number of chosen equivalents, the extent to which informative comments are added, etc. This may cause a certain asymmetry between the TLs but it is not necessarily a problem as each of the six dictionaries is independent of the others.

4. The Dictionary Contents and Infrastructure

ISLEX runs on Linux Operating System and uses Postgres relational database. All its software is written in Perl. Information is extracted from the database by SQL-queries, or sometimes by special export mechanisms that are easier to use.

It is a medium sized dictionary, with 50,000 headwords (lemmas). Many different types of data or information categories have been defined in the database. This chapter describes some of the central components and their role in the dictionary, and briefly discusses the dictionary's editorial environment.

2 Only one truly monolingual Icelandic dictionary existed then, which had been revised and reprinted several times.There are plans to publish a new monolingual dictionary on the web in the near future, based on the ISLEX data.

4.1. Contents

Lemma and word class. The emphasis is on modern Icelandic but the lemmalist also contains some important older vocabulary.

Target languages: equivalents or explanations. The non-Icelandic part of the dictionary: Danish, Norwegian, Swedish, Faroese and Finnish.

Meanings or explanations in Icelandic. These were originally included to assist the work on the TLs and are not visible to the user, partly because they were made quickly and are therefore below acceptable dictionary standard, in some cases.

Grammatical information. This includes case government of the Icelandic verbs and prepositions, among other things. Some grammatical information on the TLs is provided too, e.g. the plural of nouns and past tense of verbs. The different TLs have different solutions for how to display this, but in all instances the morphology is shown by a mouseover of the equivalent. This feature is currently available for Danish and Swedish only.

Figure 2: The inflectional forms of the Swedish word *fjäril* 'butterfly' become visible with mouseover.

Pronunciation. Pronunciations of the Icelandic lemmas are given as recorded, spoken sounds.

Pictures. ISLEX contains 3000 pictures. They are visually pleasing and give the dictionary a more relaxed, less serious feeling. Besides, they may attract younger users to the dictionary.

Phrases, fixed expressions, examples of use. ISLEX places emphasis on all sorts of phrases and collocations, and the dictionary contains a large number of them. The phraseology is fully translated into the TLs, as are the numerous examples of use.

Cross references and external links. One of the many benefits of a web-based dictionary is that clicking on a cross reference takes you there immediately. The same applies to external hyperlinks. The external links are currently mainly used to connect a lemma to its inflectional paradigm in another database[3], as seen in figure 3. There are plans to link the lemmas and phrases to a tagged corpus of Icelandic too, in a similar way as is done in *Den danske ordbog* (The Danish Dictionary) at DSL in Copenhagen[4]

Eintala			Fleirtala		
	án greinis	með greini		án greinis	með greini
Nf.	**sandur**	**sandurinn**	Nf.	**sandar**	**sandarnir**
Þf.	**sand**	**sandinn**	Þf.	**sanda**	**sandana**
Þgf.	**sandi**	**sandinum**	Þgf.	**söndum**	**söndunum**
Ef.	**sands**	**sandsins**	Ef.	**sanda**	**sandanna**

Figure 3: The inflection of the Icelandic noun *sandur* 'sand'. It has four cases, singular and plural, and indefinite and definite forms.

Semantic fields. The whole vocabulary has been divided into 650 semantic fields. To name a few, they include *animals, sound, wind, death, law, chemistry* (mainly nouns), and *nationality, pleasure, health, shape, colour* (mainly adjectives), and so on. Sometimes the fields are very fine-grained, and possibly there are too many of them. A certain Swedish study of 'domain determination' used only about 100 different labels (Malmgren & Sjögreen, 2004). However, the exactness of the semantic labels would in any case depend on the intended uses of such classification.

4.2. Editing

In the editing page, a member of the editorial team selects the relevant item of data from a drop-down list. For added clarity the different languages have differently coloured background. The order of the items is fixed: Danish at the top, then two types of Norwegian, Swedish, Faroese, and Finnish at the bottom. The morphology of Danish and Swedish has been automatically added to the equivalents by a special import mechanism, but sometimes it has been necessary to correct it manually, in particular where there are homonyms in the TLs.

Figure 4 shows the article *sandur* 'sand' in the editorial environment, a relatively simple lemma. The dictionary has a tree structure in the database, where each article is (or can be) hierarchically structured. This means that every item can have sub-items which again can have sub-items. Figure 4 illustrates this. The noun *sandur* has two different meanings. Each of these is contained in a section preceded by a number, 1 or 2. In each section the TLs are placed within the numbered item, in which there may also be a phrase and/or an example of use. On a still lower level other information can be included, such as comments related to individual items of the text. The TLs' morphology is always set one step below its parent word, as seen in two places in figure 4 (headed by MORFO-DA and MORFO-SE).

3 http://bin.arnastofnun.is.
4 http://ordnet.dk/ddo/.

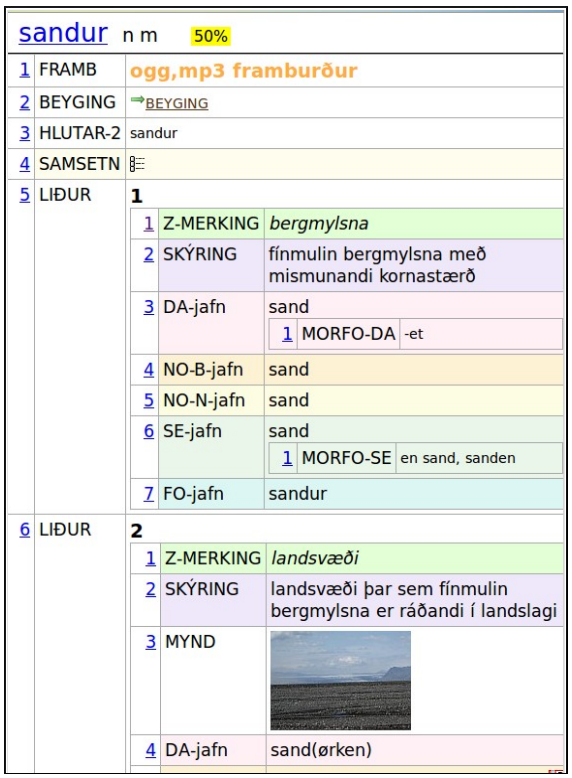

Figure 4: The editorial environment of the dictionary. The hierarchy is inherited to the user interface where different levels are shown by stages of indentation. The '50%' in the top line indicates that this lemma belongs to the more important half of the vocabulary, where the criterion used is normally the frequency of the words.

A datatype called OSAMB (short for *orðasamband* 'phrase, idiom') is shown in figure 5, in edit mode. It is a screen shot cut out of the entry *motta* ('mat, rug'). It reads *halda sig/sér á mottunni'* (literally 'keep oneself on the mat', i.e. 'restrain oneself'). This item is a parent to lines 1-8 beneath it: first an explanation in Icelandic (invisible to the user) and all the TLs in lines 2-8 in their predetermined order.

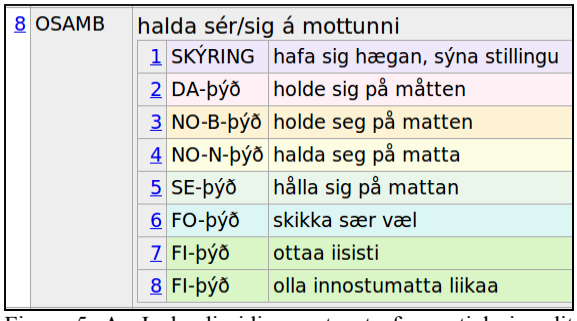

Figure 5: An Icelandic idiom cut out of an article in edit mode. In the line numbered 1 (SKÝRING) there is an explanation in Icelandic, and the seven Scandinavian translations in lines 2-8.

5. User Interface

The design of the interface of a web dictionary and its user profiles is of great importance and must be well thought out. The interfaces of multilingual dictionaries are potentially more complicated (and more confusing) than those of bilingual dictionaries, and they are not always very successfully executed. Obviously, the user must have an easy access to the lexicographic data he or she needs, and the way should be clear. Danish studies suggests that dictionary users are not very patient with complicated options. They have a tendency to use the default settings only, and the majority of users do not bother with the advanced search possibilities that lie beneath the surface. (Trap-Jensen 2010; Lorentzen & Theilgaard, 2012).

Originally ISLEX' user interface was rather more complex that it is now, but a consultation with an expert at DSL in Copenhagen resulted in a simpler surface structure and easier search functionalities.

The use of the dictionary is relatively straightforward. The search combines the Icelandic lemmas' base form, their other inflectional forms, and the TLs' equivalents. This functionality can doubtless still be improved, cf. Sanders' critical assessment (2013). The ISLEX user interface also offers advanced search with more options, but this feature is currently being revised.

Figure 6: The simple search form of ISLEX. The metalanguage is Swedish. The search combines the Icelandic lemmas' base form, their other inflectional forms, and the TLs' equivalents.

The metalanguage of the web page is set by the user (just below the landscape in fig. 6). The current options are five Scandinavian languages. The dictionary is selected as shown in figure 7. It says (in Swedish) *välj ordbok* ('choose dictionary'), followed by the four flags that represent the TLs currently available. When the Faroese and Finnish dictionaries will be ready to be opened, two more metalanguages will be added, and two more flags as in figure 7.

Figure 7: By clicking on the flags, a dictionary is selected. The options are Danish, Norwegian Bokmal, Norwegian Nynorsk, Swedish and alla 'all'. By clicking on the last option, all the TLs are shown simultaneously, as in fig. 1. This is not recommended for larger entries as they may look seriously cluttered.

In figure 8, the word *sandur* ('sand') has been looked up in the Icelandic-Swedish dictionary. The metalanguage of the web page is independent of the chosen dictionary, but here it is also Swedish.

Figure 8: The word *sandur* ('sand') in the Icelandic-Swedish dictionary. The metalanguage can be seen at the top: by how the word class is presented (*mask.*, according to the Swedish convention) and by the Swedish words *uttal* ('pronunciation') and *böjning* ('inflection'). There are two numbered senses in this article. Sense 2 ('sand desert') has a picture, taken by the author.

6. Further Work, More Uses

A project like ISLEX has many possibilities for further work. Obviously, more languages can be added as TLs. As a dictionary there are already two spin-offs of the Nordic project, firstly an Icelandic-French bilingual dictionary which began in early 2014. This dictionary will probably be bidirectional (also French-Icelandic) and in that respect different from the Nordic work. Secondly, there is a new, monolingual Icelandic dictionary under construction that has been generated from the work on ISLEX. Moreover, an experiment has been made of reversing the data between the Icelandic SL and the two Norwegian TLs, and such work may be developed further in the future.

As the work progressed through the years, it became apparent that the contents of the dictionary could be an important contribution to some language technology projects involving the Nordic languages. The ISLEX data is technically a big, largely handmade compilation of various linguistic information and a reliable source of structured, multilingual data.

Some key parts of ISLEX have already been delivered to the Icelandic repository of the METASHARE initiative (http://malfong.is), e.g. all the headwords plus their equivalents in Danish, Norwegian and Swedish. The format of the data is LMF (Lexical Markup Framework), a category of XML. The sound files of the headwords' recorded pronunciations have also been made available as open source material in three different formats (mp3, ogg and wav).

7. Conclusion

The original objective of the project was to fill a need for a dictionary between Icelandic and the other Nordic languages. Since its opening on the web in 2011, ISLEX has been extremely well received by its users. Experience has shown that the dictionary is very useful in (obligatory) Danish classes in Icelandic schools, and also in universities in the countries where Icelandic is taught. Many translators in Scandinavia are its faithful users.

Despite the dictionary's success there is still room for improvement, and a user survey conducted in March 2014 has shown that more work has to be done to better fulfill the needs of its users.

The aim of the Department of Lexicography at the Árni Magnússon Institute is, firstly, to make the dictionary more user friendly, and second, to add more target languages. The new languages will probably not become a part of the ISLEX project itself, but of its sister dictionary. However, the funding of such work can be difficult and may take some years. Until then the users are mainly restricted to the Scandinavian language zone.

ISLEX is accessible at the following web addresses: www.islex.is, www.islex.dk, www.islex.no and www.islex.se.

8. References

Beygingarlýsing íslensks nútímamáls. [The Database of Modern Icelandic Inflections.] (n.d.). Kristín Bjarnadóttir, editor. The Árni Magnússon Institute for Icelandic Studies. http://bin.arnastofnun.is.

Den Danske Ordbog. [The Danish Dictionary.] (n.d.). Society for Danish Language and Literature. http://ordnet.dk/ddo/.

ISLEX-orðabókin. [The ISLEX Dictionary.] (n.d.). Thórdís Úlfarsdóttir, editor-in-chief. The Árni Magnússon Institute for Icelandic Studies.

www.islex.is, www.islex.dk, www.islex.no and www.islex.se.

Jónsdóttir, H. and Úlfarsdóttir, Th. (2011). ISLEX – en flersproget nordisk ordbog. *Nordiska Studier i Lexikografi 11,* pp. 353-366.

Lorentzen, H. and Theilgaard, L. (2012). In *Proceedings of the XV Euralex International Congress,* pp. 654-660.

Malmgren, S-G., Sjögreen, C. (2004). Using a lexical database for domain determination, partial disambiguation and dictionary expansion. In *Proceedings of the XI Euralex International Congress,* pp. 1133-1143.

Sanders, Christopher (2005). Bilingual Dictionaries of Icelandic: Types of Users and their Different Needs - a Discussion. *Orð og tunga 7,* pp. 41-57.

Sanders, Christopher (2013). ISLEX foråret 2013. *LexicoNordica 20,* pp. 259-277.

Sigurðardóttir, A., Hannesdóttir, A., Jónsdóttir, H., Jansson, H., Trap-Jensen, L., Úlfarsdóttir, Th. (2008). ISLEX - An Icelandic-Scandinavian Multilingual Online Dictionary. In *Proceedings of the XIII Euralex International Congress,* pp. 779-789.

Sigurðardóttir, A., Hannesdóttir, A., Rauset, M. (2011). En-, två- eller flerspråkig ordbok? *Nordiska Studier i Lexikografi 11,* pp. 512-523.

Trap-Jensen, Lars. (2010). One, Two, Many: Customization and User Profiles in Internet Dictionaries. In *Proceedings of the XIV Euralex International Congress,* pp. 1133-1140.

Úlfarsdóttir, Thórdís (2013). ISLEX – norræn margmála orðabók. *Orð og tunga 15,* pp. 41-71.