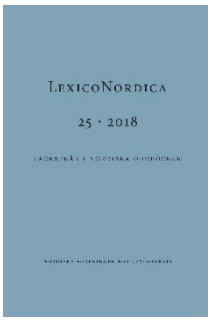
LexicoNordica

Forfatter:	Eiríkur Rögnvaldsson [Íslenskt orðanet: a treasure for writers and word lovers]
Anmeldt værk:	Íslenskt orðanet. Author: Jón Hilmar Jónsson. Design and programming: Bjarki Karlsson. Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík 2016. Online edition: <http://ordanet.arnastofnun.is/>.
Kilde:	LexicoNordica 25, 2018, s. 313-328
URL:	https://tidsskrift.dk/index.php/lexn/issue/archive



Íslenskt orðanet: a treasure for writers and word lovers

Eiríkur Rögnvaldsson

Íslenskt orðanet. Author: Jón Hilmar Jónsson. Design and programming: Bjarki Karlsson. Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík 2016. Online edition: <http://ordanet.arnastofnun.is/>.

1. The work

Íslenskt orðanet ('Icelandic wordnet'; henceforth ÍNET) is an online dictionary, or thesaurus, or database – it is not easy to classify this great work. It is amazing that it is essentially the work of one man – Jón Hilmar Jónsson, research professor (now emeritus) at the Árni Magnússon Institute for Icelandic Studies (formerly of the Institute of Lexicography at the University of Iceland). ÍNET is based on three previous dictionaries by Jón Hilmar: *Orðastaður* ('The place of words', 1994), a dictionary of collocations and multiword expressions; *Orðaheimur* ('The world of words', 2002), a conceptual dictionary; and *Stóra orðabókin um íslenska málnotkun* ('The big dictionary of Icelandic language usage', 2005), which is a combination and expansion of the two previous works.

ÍNET has been greatly expanded by harvesting examples from corpora, especially Tímarit.is, which is a digitized corpus containing the bulk of Icelandic newspapers and magazines from the beginning of the 19th century to the present, and Mörkuð íslensk málheild (MIM), which is a 25 million word balanced tagged corpus containing text samples of various genres from the first dec-

ade of the 21st century. ÍNET also benefits greatly from a large list of fixed phrases and collocations which the author has excerpted from the collections of the Institute of Lexicography. The material in ÍNET is immense – almost 200,000 headwords (102,000 single words and 93,000 phrases in April 2018 according to information on the project web).

Despite the name, ÍNET bears little relation to the well-known Princeton WordNet and similar works that have been developed for many languages in recent years. WordNet has a hierarchical structure and is a complex network of synonyms, antonyms, hypernyms, hyponyms, meronyms, holonyms, sister terms, derivationally related forms, etc. Every word belongs to a synset, a set of cognitive synonyms. In contrast, ÍNET has a flat structure where semantically related headwords are grouped together under a concept, but the concepts themselves are not directly related. Individual words can appear under more than one concept, but many words have not been connected to any concept at all. It must be emphasized that even though ÍNET was officially opened in 2016, it is still a work in progress and the semantic classification is still ongoing.

It is impossible to do justice to such a voluminous work in a short review, but let me start by showing the opening screen of the website.

Figure 1: The opening screen of Íslenskt orðanet.

Under the icon in the upper right corner we can find basic information about the project – its main characteristics, history, statistics, and people. This information, and the user interface in general, is only in Icelandic. Under the question mark to the right we find search help, and context-sensitive help is available for all the different screens on the website.

We start by entering a word in the search box (*leita* 'search') and press "enter" or the arrow to the right of the search box. If we enter the word *regn* 'rain', for instance, we get the screen shown in figure 2.

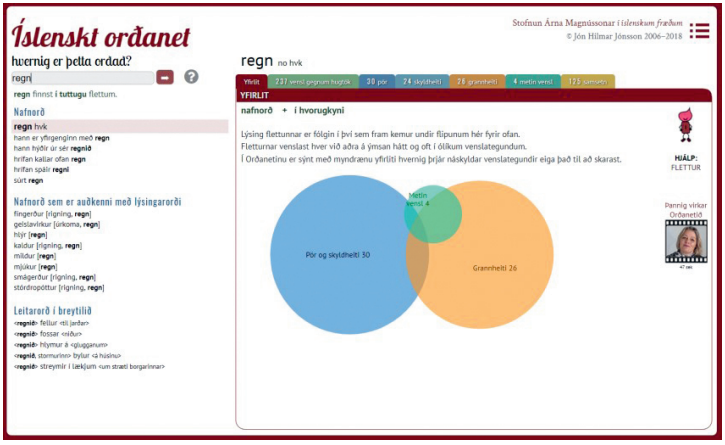


Figure 2: The main screen of Íslenskt orðanet.

The screen is divided into two parts – a search window to the left, taking up approximately one-third of the screen, and a results window to the right, taking up the rest of the screen. Let us now look at these windows in turn.

2. The search window

When a word is entered into the search box, the word itself appears below (provided it is a headword in the database) followed by a list of phrases containing it. This list can contain from zero to several hundred phrases (for instance, 1585 for the verb *gera* 'do', 481 for the adjective *góður* 'good' and 365 for the noun *barn* 'child').

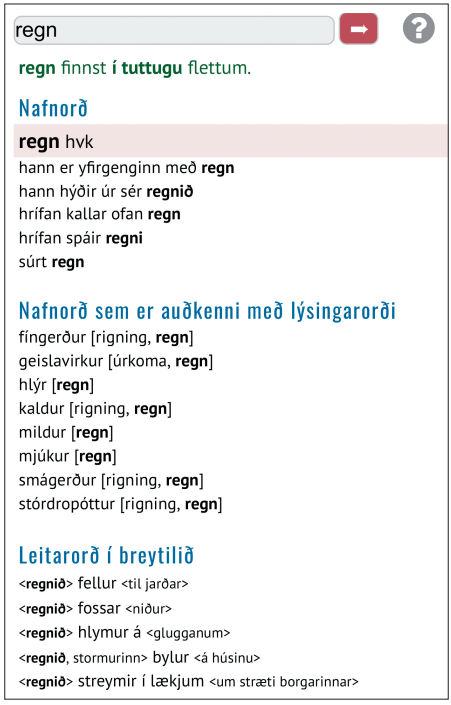


Figure 3: The search window.

The list is usually divided into two or three sublists with separate headings, and the criteria for this division are a bit different for different parts of speech. In figure 3, we see the search results for *regn*. The list starts with this word and typical phrases containing

it, under the heading *Nafnorð* 'noun'. The next section of the list shows adjectives that are characteristic as attributes to the noun in question, such as *fíngerður* 'fine', *geislavirkur* 'radioactive', *mjúkur* 'soft', etc. The final section lists phrases with variables where the search word is a typical – and in some cases the only possible – way to complete the phrase, such as *<regnið> fossar <niður>* '<the rain> pours <down>', *<regnið> hlymur <á húsinu>* '<the rain> thunders on <the house>'.

If the user clicks on one of these phrases, it jumps into the search box and information on that phrase appears in the results window to the right, instead of information on the original search word.

The variables exhibit some inconsistencies, especially with respect to genders. Many of them show both a masculine and feminine personal pronoun, such as *hjálpa <honum, henni> á fætur* 'help <him/her> on [his/her] feet', whereas others only show one pronoun (usually masculine), such as *vera fús að <hjálpa honum>* 'be willing to <help him>'. There does not seem to be any general rule as to whether both possibilities are shown.

Given the great number of headwords, one would expect ÍNET to cover modern Icelandic vocabulary very well. Since it is not possible to see an alphabetical list of headwords, it is difficult to find out whether any words that should have been listed are missing. In repeated searches, however, I came across a few such instances. Neither *göltur* 'boar' nor *gylta* 'sow' is found, although both *svín* 'pig' and *grís* 'piglet' are. Some verbs are not listed in their own right, so to speak, but only as parts of phrases. Thus, a search for *lofa* 'promise' returns *lofa betrun* 'promise to behave better', and a search for *mæla*, which is actually the infinitive of two verbs which conjugate differently, 'speak' and 'measure', returns *mæla bert* 'speak openly'.

Under most or all adjectives *X*, the phrase *vera X* 'be X' is also listed as a search term – and the same goes for many nouns which

denote human characteristics. Thus, *vera þreyttur* 'be tired' is listed under þreyttur 'tired', *vera gáfaður* 'be bright' is listed under *gáfaður* 'bright', *vera aumingi* 'be a loser' is listed under *aumingi* 'loser', and so on. This may of course be justified if *vera X* has a special meaning, not fully predictable from the meaning of *X*, such as in *vera grænn*, which literally translates to 'be green' but usually means 'naive' when used about a person. In most cases, however, separate listing of *vera X* appears to be superfluous – and can be misleading since analogous pairs can appear under different headwords, as explained below.

3. The results window

In the results window, information on the search word is displayed under different tabs, from one up to six, each with its own distinctive colour. None of these tabs is obligatory, but the six shown in figure 4 are typical for individual words.

The leftmost tab is *vensl gegnum hugtök* 'relations through concepts', the next one is *pör* '(word) pairs', the third tab is *skyldheiti* 'related words', then comes *grannheiti* 'neighbouring words', the fifth tab is *metin vensl* 'judged relations', and finally we have *samsetn(ingar)* 'compounds'. If the search term is a phrase, *setningargerð* 'syntactic structure' appears as the rightmost tab.

In the lower half of the results window, the number of different relations, the ratio among them, and their overlap is shown using circles of different sizes and colours.

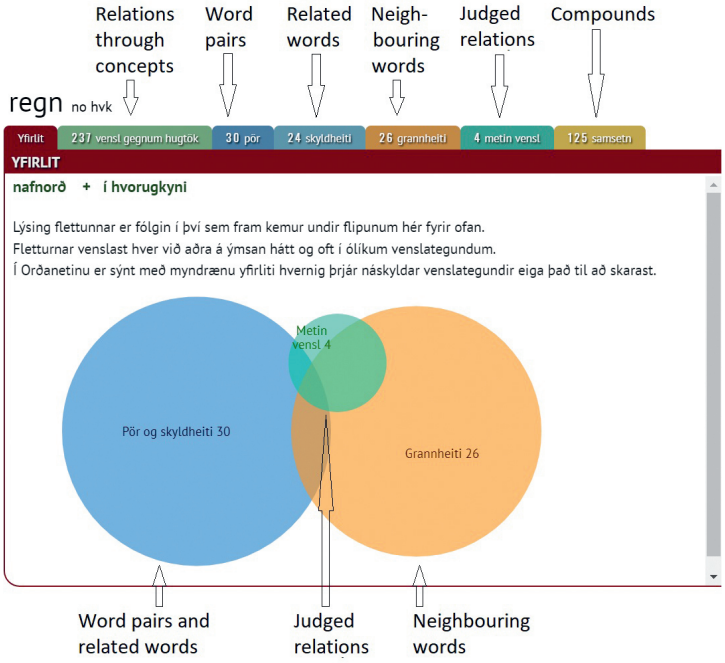


Figure 4: The results window.

3.1. Concepts

Under "relations through concepts" we get an alphabetized list of all individual words and phrases that have been classified under the same concept as the word we have searched for. The search word *regn* 'rain' is classified under only one concept, RIGNING 'rain'. In many cases, however, the search word has been classified under two or more concepts. Both (or all) will then appear under this heading, and the user can choose among them to see the relevant related words. Another column under "relations through concepts" has the heading "related headwords through pairs". The words in this column occur in pairs with words in the first col-

umn. The third column lists concepts related to the selected concept heading the first column. The closeness of the relationship is measured by the number of common words that occur in pairs related to both concepts. Thus, the relationship index between the concepts RIGNING and ÞOKA 'fog' is listed as 17, which should mean that the same 17 words can be found in pairs listed under both concepts.

The selection of concepts and the grouping of words under certain concepts appear to be rather haphazard, and I will just name a few examples of the numerous I have come across. The words ær 'ewe' and *lamb* 'lamb' are listed under BÚFÉ 'livestock', but *hrútur* 'ram' is not listed under any concept. The word *bíll* 'car' is listed under two concepts, BÍLL and FARARTÆKI 'vehicle', but its exact (but more formal) synonym *bifreið* is not listed under any concept. All colours (including *hvítur* 'white') appear to be listed under LITUR/ LITABRIGÐI 'colour / shade of colour', except *svartur* 'black', which is not listed under any concept.

The word *gluggatjöld* 'window curtains' is only listed under one concept, which happens to be GLUGGATJÖLD. The singular *gluggatjald* is also a headword, even though the word is almost never used in the singular as evidenced by the phrases related to this headword, which all show the plural. Surprisingly, however, *gluggatjald* is not only grouped under GLUGGATJÖLD but also under HÚSBÚNAÐUR 'furnishings'. The word *gardína*, which is an exact synonym of *gluggatjöld* (the only difference being that *gardína* is a loanword from Danish whereas *gluggatjöld* is a neologism) is also listed under these two concepts. Phrases containing the plural *gardínur* (which is not a headword by itself) are, however, either only listed under GLUGGATJÖLD (*hengja <gardínur> <fyrir gluggann>* 'hang up <curtains> for the window') or under no concept at all (*setja upp <gardínur>* 'put up <curtains>').

Since for many adjectives (and nouns) *X*, the phrase *vera X* 'be X' is also a headword, as mentioned above, these headwords each

have their own listing of concepts, word pairs, related words, and neighbouring words. This often leads to strange discrepancies. For instance, *vera umdeildur* 'be controversial' is listed under GAGNRÝNI 'criticism' but *umdeildur* 'controversial' is not listed under any concept. The word *greindur* 'intelligent' is listed under both GREIND/GÁFUR 'intelligence/brightness' and VIT/ SKYNSEMI 'wisdom/common sense', whereas *vera greindur* 'be intelligent' is only listed under GREIND/GÁFUR. Conversely, *breyskur* 'fallible' falls under VEIKLYNDI 'weakness', but *vera breyskur* 'be fallible' is listed under both VEIKLYNDI and BREYSKLEIKI 'fallibleness'. The word *móður* 'short of breath' falls under ÞREYTA 'tiredness' whereas *vera móður* 'be short of breath' falls under MÆÐI 'shortness of breath'.

3.2. Word pairs

Under "word pairs" we get a list of pairs where the search word is either the first or the second member – pairs like *regn og bleyta* 'rain and wetness', *regn og myrkur* 'rain and darkness', *ský og regn* 'clouds and rain', *regn og sólskin* 'rain and sunshine', *stormur og regn* 'storm and rain', *regn og þoka* 'rain and fog', etc. These pairs are actually the backbone of the work, in the sense that they are central in deducing and assessing the relations between words. The pairs are mainly taken from corpora, especially Tímarit.is and MIM, as mentioned above. There is no doubt that this use of corpora adds an invaluable dimension to the work and makes the semantic classification much more detailed and accurate than otherwise would have been possible.

Since many of the pairs seem to be taken directly and unchanged from texts, they exhibit a lot of inconsistency. Under *glaður* 'glad', for instance, 70 pairs are listed, all but one with *vera glaður* 'be glad'. In 65 of these pairs, *glaður* is in the masculine, but 4 pairs have the feminine *glöð* instead – for no obvious reason. In

many cases, the same noun is in the singular in some pairs (*hestur og asni* 'horse and donkey', *tölva og skjár* 'computer and monitor') but in the plural in others (*hestar og hundar* 'horses and dogs', *tölvur og skjávarpar* 'computers and projectors'). On the other hand, verbs always seem to be in the infinitive. This diversity should not lessen the usability of the dictionary, even though it may be a bit confusing for the user.

Word pairs with *vera X* 'be X' are either listed under the headword *vera X* or under *X*, and there does not seem to be any general rule as to their distribution. For instance, 220 pairs are listed under *gáfaður* 'bright', all of them containing *vera gáfaður* 'be bright'. These pairs could of course have been listed under *vera gáfaður*, but there only 8 pairs are found – some of them the same as are listed under *gáfaður*, but not all. 26 pairs are listed under *svangur* 'hungry', all but one with *vera svangur* 'be hungry'. No pairs are listed under *vera svangur*.

Despite the great number of pairs, I found some rather common pairs to be missing. Examples of these are *góður og gegn* 'good and honest' (1158 instances on Tímarit.is), *æpa og góla* 'scream and wail' (18 instances on Tímarit.is), and *dauði og djöfull* 'death and devil' (202 instances on Tímarit.is). It must be mentioned, however, that even though *dauði og djöfull* does not appear in the list of word pairs, neither under *dauði* nor under *djöfull*, the combination *dauðann og djöfulinn* / *dauðanum og djöflinum* 'the death and the devil ACC/DAT' occurs four times in phrases under *dauði*. The pair *brjóta og bramla* 'break and destroy' is listed as a phrase under *brjóta*, but not as a pair (only one pair is listed under *brjóta*, *brjóta og eyðileggja*, which has the same meaning as *brjóta og bramla*).

Furthermore, the pairs sometimes contain too detailed information and superfluous words. Under *dauði* we find for instance the pairs *dauði <Masaryks> og valdarán <kommúnista>* 'the death of <Masaryk> and the <communist> coup' and *dauði <Sulla> og valdataka <Cesars>* 'the death of <Sulla> and <Caesar's> taking

of power'. I do not see the reason for including the names in these cases.

3.3. Related words

The list under the tab "related words" is automatically generated, based on common words in pairs. Thus, *ský* 'cloud' is listed as related to *regn* 'rain' because two words, *vindur* 'wind' and þoka 'fog' form pairs with both of them – *vindur og regn*, *regn og þoka* vs. *ský og vindur*, *ský og þoka*. The number of common words is shown, and also their ratio of the pairs of the related word. By clicking on a magnifying glass to the left of each related word, we get a list of the common words, and the overlapping of the pairs for the search word and the related word is shown with circles of different sizes.

However, the use of pairs to determine word relationship can sometimes be a bit misleading. In some cases, it seems that the pairs are made up of words that have happened to co-occur in a text without being specially related. For instance, five pairs are listed under *fótaaðgerð* 'pedicure' – *fótaaðgerð og hárgreiðsla* 'pedicure and hairstyle', *hársnyrting og fótaaðgerðir* 'hairstyling and pedicure', *leikfimi og fótaaðgerðir* 'gymnastics and pedicure', *fótaaðgerð og myndlist* 'pedicure and art', and *fótaaðgerðir og sund* 'pedicure and swimming'. None of these words appears to be specially related to *fótaaðgerð* 'pedicure' (although *hárgreiðsla*/*hársnyrting* and *fótaaðgerð* can be claimed to belong to the same semantic field in some sense). In spite of that, both *leikfimi* and *sund* also appear under 'related words' – based on the pairs *leikfimi og fótaaðgerðir* and *fótaaðgerðir og sund*, since *sund og leikfimi* is also a pair.

Nine pairs are listed under the word (and concept) *gluggatjöld* 'window curtains'. Exactly the same nine words that occur in pairs with *gluggatjöld* are also listed under 'related words' to *gluggatjöld* even though many of them do not seem that related, such as *handklæði* 'towels', *húsgögn*, *innanstokksmunir* 'furniture', and *rúmföt*,

sængurfatnaður, *sængurföt* 'bed linen'. On the other hand, no pairs, and hence no related words, are listed under the singular *glugga-tjald*, which is also a headword as mentioned above.

I have serious doubts about the usefulness of showing word pairs and related words in the ÍNET interface. The pairs are too irregular and too accidental to be of much value for the general user, and since the "related words" are automatically derived from the pairs, they are also in many cases too accidental and not really related to the search word.

Under "related words" for *góður* 'good', for instance, we get 118 words. Among them are listed words like þurr 'dry' and *gamall* 'old', even though each word only has one pair containing a word that also occurs in a pair with *góður*. True, the number of common words is shown, together with the percentage of common words with respect to the number of pairs (*gamall* occurs in 166 pairs, and since only one of them also occurs in a pair with *góður*, the ratio is 0.6%), but I think it would be much better to have a threshold here and only show words which reach a certain limit, both with respect to the number of common words and to their ratio of the number of pairs.

3.4. Neighbouring words and judged relations

The words we find under the tabs "neighbouring words" and "judged relations" are closer to what we find in traditional synonym dictionaries. The list of "neighbouring words" is based on word pairs from texts, but also on the list of fixed phrases and collocations mentioned in section 1. This list is not automatically generated and does not just contain any words that are connected to the search word through pairs, but only those that the author has judged to be semantically close. In many cases, these words are close to being synonyms of the search word.

By clicking on an icon (chain links) to the left of the search

word, we get a list of the pairs and phrases on which the relation rests. Thus, the words *regn* 'rain' and *skúr* 'shower' are related through the words *leysing* 'melting (snow)' (*með regni og leysingu*, *leysing og skúrar*, from the list of pairs) and *þéttur* 'dense' (*þétt regn*, *þéttur skúr*, from the list of fixed phrases of collocations).

In contrast to the other relations, "judged relations" are not based on text examples, but only on the subjective judgement of the author. In most cases, these relations are synonyms, but in the case of adjectives, antonyms may also be listed. The third possibility under this tab is *stikla* 'tip' which is used for words that are related to and relevant to the search word, but do not show up in connection to it in the data.

I find the information under these two tabs very useful. The extensive use of examples from corpora adds a new dimension to the traditional synonym dictionary and brings the dictionary more up to date.

3.5. Compounds and syntactic structures

The rightmost tab for individual words is "compounds". Under that tab we get two separate lists of compounds – one where the search word is the first immediate constituent of the compound and another where it is the second constituent. The other part of the compounds is usually also a headword, but not always, and the same goes for the compounds themselves. For instance, many compounds with *regn* 'rain' like *regn-frakki* 'raincoat' and *sprengju-regn* 'bomb rain' are headwords, whereas *regn-lækur* 'rain brook' and *haust-regn* 'autumn rain' are not, even though all are listed under the "compounds" tab. Notice that separate searches for *regn-lækur* and *haustregn* do not give any hits, although these words clearly exist in the database. Thus, it might be useful to point out to users to look under simple words for compounds not found with direct search.

If the search term is a phrase, its syntactic structure is described under a separate tab. To name an example, *<regnið> fossar <niður>* '<the rain> pours <down>' is described as *<nafnorð í nefnifalli með greini> » sagnorð » <atviksorð/atviksliður>* '<noun in the nominative with a definite article> » verb » <adverb/adverbial phrase>'. All phrases having identical syntactic structure, regardless of the words they contain, are then listed below the description – and there can be hundreds of them. I must admit that I do not see the usefulness of this list.

4. Design, interface, and search

The usefulness and success of an online dictionary depends heavily on the design and user-friendliness of the interface, and on the flexibility and efficiency of the search. ÍNET scores high on these criteria. The user interface and the graphic design is for the most part well-conceived and serves its purpose. I used Google Chrome running on Windows 10 for this review. ÍNET also works fine on other browsers I tried (Internet Explorer, Mozilla Firefox, Vivaldi). No special mobile interface appears to be available, and using ÍNET on mobile phones is therefore a bit cumbersome.

The search is very flexible and extremely quick. Many of the lists can be ordered according to different criteria, such as part of speech, number of relations, ratio of relations, the alphabet, etc. This is very convenient and highlights the advantages of an online dictionary compared to traditional printed dictionaries.

However, I have my doubts about the usefulness of the circles that show the ratio of neighbouring words and judged relations versus word pairs, and the overlapping of these categories. The graphic illustration of the relationship between neighbouring words is also not very attractive.

I have encountered a few technical flaws in ÍNET (which may

have been fixed by the time this is published). For instance, the linking seems to have gone wrong in some cases. To name just two examples: In the list of pairs for *frægur* 'famous', the pair *frægt og umdeilt <skáldrit>* 'famous and controversial <work of fiction>' appears twice. If we click on the first instance, we get the adjective *umdeildur*, but if we click on the second instance, we get (correctly) the phrase *umdeildur [ráðstöfun; ákvörðun; mál, rit]* 'controversial [operation; decision; case, book]'. An exactly parallel case is found in the list of pairs for *góður* 'good' where *góð og uppbyggileg <umræða; boðskapur>* 'good and inspiring <discussion; message>' appears twice.

Finally, I can mention that it is not always possible to use the back button in the browser – sometimes it sends the user to the initial screen of the website instead of the previous page.

5. Use – and potential usefulness

Despite certain shortcomings and inconsistencies, ÍNET is an invaluable tool for the ordinary user in its current online version. It should be especially useful for writers, journalists, translators, and others who want to write good and varied text using rich vocabulary. However, even though the interface and design of ÍNET is user-friendly as pointed out above, and even though context-sensitive help is available for all screens, the fact remains that this is a complex work which opens up many possibilities and it takes some time to familiarise oneself with it and figure out how to make the most of it. My main concern is that many potential users will not be patient enough.

I am not a lexicographer but a linguist who has worked extensively with corpus linguistics and language technology and this inevitably affects my viewpoint. I think the usefulness of ÍNET would be greatest within language technology. Its wealth of in-

formation on word combinations and relations would be an extremely valuable resource for current methodologies like neural networks and would facilitate disambiguation, genre classification, machine translation, style checking etc. I believe that the source files of ÍNET are potentially the most valuable resource that exists for enhancing Icelandic language technology, and it is to be hoped that they will be made available for that purpose in the future.

References

Jón Hilmar Jónsson (1994): *Orðastaður. Orðabók um íslenska málnotkun.* Reykjavík: Mál og menning. [2nd edition 2001. Reykjavík: JPV.]

Jón Hilmar Jónsson (2002): *Orðaheimur. Íslensk hugtakaorðabók*. Reykjavík: JPV.

Jón Hilmar Jónsson (2005): *Stóra orðabókin um íslenska málnotkun.* Reykjavík: JPV.

MIM = Mörkuð íslensk málheild. <http://mim.arnastofnun.is> (October 2018).

Tímarit.is = <http://timarit.is> (October 2018).

WordNet = <https://wordnet.princeton.edu/> (October 2018).

Eiríkur Rögnvaldsson
professor emeritus
Háskóla Íslands
Árnagarði við Suðurgötu
IS-101 Reykjavík
eirikur@hi.is