NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Breaking away from tradition: Linking a database of inflection to an electronic dictionary

Forfatter: Kristín Bjarnadóttir

Kilde: Nordiska Studier i Lexikografi 11, 2012, s. 129-137
Rapport från Konferens om lexikografi i Norden, Lund 24.-27. maj 2011

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive

Breaking away from tradition: Linking a database of inflection to an electronic dictionary

Kristín Bjarnadóttir

Icelandic inflection is complex and there is an abundance of variant forms, which are very often changeable according to time and style. The choice of variants within a paradigm in some cases depends on meaning and context. The tradition in Icelandic dictionaries is to give the principal parts of words, as indicators of inflectional classes. These, however, contain insufficient information for the remainder of the paradigm to be predictable. In an online dictionary this problem can be solved by a link to a database of inflection, such as the Database of Modern Icelandic Inflection (DMII). This is the method chosen in ISLEX, the Icelandic-Scandinavian online dictionary, and as a result the information provided far exceeds the tradition in Icelandic dictionaries. Both projects are under construction at The Árni Magnússon Institute for Icelandic Studies.

The breakaway from tradition also takes place in the actual production of the paradigms for the DMII. Icelandic textbooks and handbooks on morphology describe the system as a whole, i.e., they are constructed 'top-down', with a limited number of examples for each inflectional class. The production of DMII is, in contrast, 'bottom-up', as it is based on research on individual words, not inflectional classes. The aim is to provide enough information for the user to make an educated choice by himself.

1. Introduction

The topic of this paper is the link-up of two projects in progress at The Árni Magnússon Institute for Icelandic Studies in Reykjavík, i.e. the web-based Icelandic/Danish, Norwegian and Swedish dictionary ISLEX,[1] and The Database of Modern Icelandic Inflection, BÍN (Beygingarlýsing íslensks nútímamáls).[2]

---

1 ISLEX. Halldóra Jónsdóttir, Þórdís Úlfarsdóttir (red.) The Árni Magnússon Institute for Icelandic Studies, Reykjavík. http://islex.hi.is; islex.dk; islex.no; islex.se.

2 BÍN = Beygingarlýsing íslensks nútímamáls. Kristín Bjarnadóttir (red). The Árni Magnússon Institute for Icelandic Studies, Reykjavík. http://bin.arnastofnun.is/.

Instead of the traditional information on inflection in Icelandic dictionaries, each headword in ISLEX is connected to a full paradigm for the word in BÍN by a hyperlink. This provides the user with much more information on inflection than hiterto possible in Icelandic dictionaries. The information on inflection contained in BÍN can also be said to break away from the Icelandic tradition of treatment of the inflectional system by using a 'bottom up' approach to describe the system, instead of the 'top-down' approach traditionally used in grammatical description. The difference is that of a survey of the system as a whole, as demonstrated in textbooks, in contrast with showing the paradigms for every single word, showing the inflection "as is", i.e. actual usage, as attempted in BÍN. It must be stressed, however, that work on BÍN is still an ongoing process, and a surprizing aspect of the work has in fact been how deficient the existing research material is when it comes to describing the inflectional system of Icelandic in the necessary detail for this approach.

2. The two projects, ISLEX and BÍN

ISLEX is an Icelandic/Danish, Norwegian, Swedish online dictionary being prepared at The Árni Magnússon Institute for Icelandic Studies, in cooperation with DSL (Det Danske Sprog- og Litteraturselskab) in Copenhagen, The University of Bergen, and The University of Gothenburg (Halldóra Jónsdóttir & Þórdís Úlfarsdóttir, this issue). ÍSLEX contains 50,000 entries. The work commenced in 2006, and ISLEX is due to open on November 16th 2011.[3]

Work on The Database of Icelandic Inflection (BÍN) commenced in 2002 at The Institute of Lexicography (Orðabók Háskólans), now the Department of Lexicography of The Árni Magnússon Institute for Icelandic Studies. BÍN was initially created to serve two purposes, i.e. to produce data for use in language technology projects, and to make the results available to the general public online. From the outset, the intention was also to utilize the data in lexicography. Work on the project started in 2002, as a part of an language technology program launched by the Minister of Education, Science and Culture (Rögnvaldsson et al., 2009). From the beginning, the aim was to show "all" and only existing word forms from the modern language in a set of paradigms. At present BÍN contains 270,000 paradigms, 5.8 million inflectional forms.

[3] ISLEX was opened as planned. Work on Faroese is now in progress, and work on Finnish is expected to begin in 2012.

3. Information on Inflection in Icelandic Dictionaries and Grammars

The tradition in Icelandic dictionaries is to give partial information on inflection, i.e. the principle parts of words. These are not necessarily sufficient to predict all the remaining inflectional forms of the word. The principle parts are the entry forms, followed by the the genitive singular and the nominative plural for nouns, and the past tense and past participle for verbs. (For weak verbs the past tense form is in the singular, for strong verbs the past tense is singular and plural. The past participle is omitted when it is predictable from the past tense. In the case of umlaut, the subjunctive past tense is added.) The inflection of adjectives is only shown in irregular cases. The examples here are from *Íslensk orðabók* (Mörður Árnason, 2007):

sokkur kk *-s, -ar*	'sock' masc., gen.sg. *sokks*, nom.pl. *sokkar*
þröskuldur kk *-s/-ar, -ar/-ir*	'threshold' masc., gen.sing. *þröskulds/þröskuldar,* nom.pl. *þröskuldar/þröskuldir*
labba s. *-aði*	'walk' v., p.t. *labbaði*
biðja s. *bað, báðum, beðið;* vh.þt. *bæði*	'pray' v., p.t.sg. *bað*, p.t.pl. báðum, p.ptc. *beðið;* subj.p.t. *bæði*
góður l. (mst. *betri,* hst. *bestur*)	'good' adj., comp. *betri*, sup. *bestur*
reiður l.	'angry' adj. [No information on inflection.]

Table 1. Examples of information on inflection in Íslensk orðabók (2007).

Until the arrival of electronic editions, dictionaries have, by reasons of space, been constrained by their format to condense inflectional information to the barest minimum, as evidenced by the examples above. As the principle parts shown are not always sufficient to predict the whole paradigm of words, this sometimes leaves the user stranded, or at best with grammar books as the only solution. As the function of the grammar books is to give surveys of the system, not to give information on individual words, the user can still be stranded, i.e. if he happens to be searching for words that happen not to be among the relatively few examples selected for demonstration in the grammar books. Furthermore, some parts of the system are systematically missing from the literature, as will be demonstrated later.

4. The BÍN Paradigms

Morphologically, Icelandic is a rich language, with 16 inflectional forms to a noun (4 cases, singular and plural, +/-definite), 120 inflectional forms to an adjective (3 genders, 4 cases, singular and plural, +/-definite; positive, comparative, superlative), and 107 inflectional forms to a verb, (indicative/subjunctive, present/past, person, number, etc.). Inflectional form here signifies a word form with grammatical tag. These figures are for a full paradigm, excluding variants. These variants are the result of the fact that the inflectional endings that mark grammatical categories are not unique, and the result is a proliferation of dual forms. In such cases, the tradition in Icelandic grammar books is to say that a word can belong to two inflectional classes. In BÍN, however, each headword or lemma is shown in full in one paradigm, i.e. a word belongs to one inflectional class only. Each inflectional class therefore includes all variants, and the result is a proliferation of inflectional classes, or over 600, as of summer 2011. The inflectional classes are only a tool for the production of the paradigms, and the users of BÍN and ISLEX do not have access to the classification. The only visible access is the individual paradigm, represented in the traditional manner, as shown in the paradigm for *sokkur*.[4]

sokkur

Karlkynsnafnorð

Eintala			**Fleirtala**		
	án greinis	með greini		án greinis	með greini
Nf.	sokkur	sokkurinn	Nf.	sokkar	sokkarnir
Þf.	sokk	sokkinn	Þf.	sokka	sokkana
Þgf.	sokk	sokknum	Þgf.	sokkum	sokkunum
Ef.	sokks	sokksins	Ef.	sokka	sokkanna

[4] The metalanguage for BÍN is as yet only in Icelandic. *Karlkynsnaforð*: n.masc., *eintala:* sg., *fleirtala:* pl., *án greinis:* indefinite, *með greini:* definite, *Nf.:* nominative, *Þf.*: accusative, *Þgf.*: dative, *Ef.*: genitive.

The link-up between ISLEX and BÍN is done simply by placing hyperlinks in each entry in ISLEX, giving the users of ISLEX the same access to BÍN as other users of the website.[5]

5. Variants in BÍN

In BÍN, an attempt is made to order variants according to use, i.e. by placing the most frequent, the most acceptable one, etc. before those of lesser importance. This is by no means easy to do, as there are many criteria to be taken into account. For technical reasons, there is, however, no way of marking the variants inside the paradigm in any manner, such as by font or colour. BÍN was initially created mainly for language technology use, where the database has to be as inclusive as possible, but the ordering of variants is not of prime importance, at least not in the language technology projects contemplated so far. The reason for inclusion is obvious; if a word form is not in the database, it will not be included in the output used for a seach engine, etc. Speakers of the language, on the other hand, want information on the use of the variants, not just a list of them. Therefore, the online version of BÍN contains notes on the choice of variants, usage, etc., which are included at the top of the paradigm. These notes are intended to help the user pick the right variant, or at least to keep him from using highly restricted ones, stylistically or otherwise, as seen in an English translation of the note on the noun *rödd* below:

Note: Sometimes an obsolete inflectional form, *röddu*, appears in the dative singular in texts:

> *Þeir hrópuðu hárri röddu.*
> 'They shouted in a loud voice.'

The same word form may very rarely appear in the accusative singular:

> *Þeir heyra ekki röddu hans.*
> 'They do not hear his voice.'

Table 2. The note on the use of variants in the noun rödd *'voice'*

The notes in BÍN are gradually being expanded, in part in response to online users demanding help with variants. Although the subject matter of BÍN is Ice-

[5] The data from BÍN is also used in the ISLEX search engine, giving users access to entries through the use of any inflectional form, as in *kött, ketti, kattar, kötturinn, köttinn, kettinum, kattarins, kettir, köttum, katta, kettirnir, kettina, köttunum, kattanna* → *köttur* 'cat'.

landic inflection in modern Icelandic, it has proved necessary to include notes on obsolete forms as these are very much a part of today's language, in idioms, fixed expressions and quotations, etc.

6. The Unpredictable Forms

In the instance of *sokkur*, the traditional dictionary representation of the inflection would be *-s, -ar*, as referred to above. This information, however, does not give the necessary information on the dative form, which is not predictable from the principle parts or from the rules found in grammar books. In fact it is not predictable in any manner, as the comparison of phonetically similar words below shows.

Masculine nouns, singular, indefinite				
nom.	*sokkur*	*flokkur*	*lokkur*	*kokkur*
acc.	*sokk*	*flokk*	*lokk*	*kokk*
dat.	***sokk***	***flokki***	***lokk/lokki***	***kokk***
gen.	*sokks*	*flokks*	*lokks*	*kokks*
'sock'	'flock, party'	'lock' (of hair)		'chef'

Table 3. The unpredictable dative of 'sokkur'.

The definite forms of the dative singular of these words are *sokknum, flokknum, lokknum, kokkinum/kokknum*, showing that the definite forms can not be predicted from the indefinite forms either, complicating the issue even further.

The dative singular of masculine nouns is notoriously unpredictable, and acknowledged in the grammatical literature as such (Friðrik Magnússon, 1984). Even so, people have a tendency to generalize, and to assume that words such as *sokkur* should have *-i* in the dative, as the grammar books suggest *-i* in the dative singular for masculine nouns ending in *-ur* in the nominative singular, apart from a few listed exceptions. This is of course the result of the fact that the grammars are surveys, intended to demonstrate how the system works, not exact manuals on the behaviour of individual words. The problem is that people will regard the rules in the grammar books as absolutes, and stick to them, at least consciously, as when answering questions on usage, even though they may never actually use the inflectional forms they consider or mark as 'correct'. This is the case with the dative of *sokkur* 'sock' cited above. The dative form *sokk* is attested in the sources at the Árni Magnússon Institute for Icelandic Studies, i.e. in the examples in the archives and text collections of 65 million running words, as well as in the available sources from Old Icelandic. The usage seems to

be unequivocal, as in: *"Ég fór úr blautum sokk/*sokki"* 'I took off a wet sock'. Search on the web gives the same result, although one example was found of the dative *sokki.*

To reach conclusions such as the one on the dative of *sokkur,* every possible source is consulted in the course of the production of the paradigms for BÍN, as far as time allows. The main sources are archives and text collections of the Árni Magnússon Institute for Icelandic Studies, and other available text collections, such as those at the University Library. The database is constantly revised.[6]

7. New Information on the Inflectional System

According to the literature, the case of the dative singular of neuter nouns should be simpler than that of the corresponding masculine ones, as the dative singular marker *-i* is supposed to be (almost) universal in the neuter. The only exceptions listed in a recent survey on Icelandic morphology are *tré* 'tree', *fé* 'money; livestock', and *hné/kné* 'knee' (Guðrún Kvaran, 2005). The *-i* is in fact the dative singular marker for single syllable and affixed words from the inherited Icelandic vocabulary, as for compounds formed from these. The research for BÍN shows, however, that neuter loanwords, especially multisyllable words, very often do not have an ending in the dative singular. Words of this type are not mentioned at all in most grammars, although some of them are by no means new in the language, such as the words *fenikel, kanel, vítríól, stúdíum, examen, fiaskó* and *ópíum* which are attested in the archives of the Árni Magnússon Institute for Icelandic Studies in citations from the turn of the 18th century.[7]

The neuter loanwords cannot be left out of BÍN, of course, although they find no place in the grammars. BÍN would be of less use in language technology if certain inflectional classes were left out, even if they exhibit traits not wholly in compliance with the historical facts on the inflectional system. Bisyllabic or multisyllabic single morpheme stems are fairly alien to the language, but their existance is a fact in the loanwords. Some of the 18th century loanwords mentioned above do in fact have Icelandic equivalents that have replaced them in the modern language. A case in point is *examen*, which has been superseded by *próf*, whereas the word *ópíum* is in full use, without a successful attempt at sub-

---

6 Users of the online version of BÍN contribute generously with comments on revision, additions and corrections. There were 398,768 visits in the period May 2010 to May 2011, 326,559 of those from Iceland.

7 Ritmálssafn Orðabókar Háskólans: http://arnastofnun.is/page/arnastofnun_gagnasafn_ritmal.

stitution by a native word. The number of these nouns in use is large enough not to make it possible to ignore them altogether and therefore they have to be a part of BÍN. That, however, makes it necessary to engage in research needed to find out how they behave in context.

The result is not as neat as that of the grammatical surveys, and here is where the difference between those and BÍN comes into focus. The surveys are 'top-down', i.e. they describe the system as a whole, choosing examples to demonstrate inflectional classes and exceptions from those. In BÍN, each paradigm stands on its own, and the inflectional classes are no more than a convenient way of producing the paradigms. When a new variant necessitates a new inflectional class, one is simply added. The system is truly 'bottom up'. Instead of a simple rule in the textbooks, stating that the dative ending of neuter nouns is *-i*, with the exceptions *tré, fé, hné/kné*, the facts on the neuter loanwords in BÍN are as follows:

The dative of neuter loanwords in BÍN: -i or -0?
- Multisyllable words can have *-0* or *-i,* sometimes depending on stem structure: *ópíum* -0, *statíf* -i
- Very many loanwords have variants: *bíó* -0/-i, *glúten* -i/-0.
- Some subgroups can never have the ending -i. This is true of the names of countries: *Íran, Ísrael, Mexíkó:*

*Ég var að koma frá Mexíkó/*Mexíkói* 'I just came from Mexico'

In contrast, names of countries that are a part of the inherited vocabulary, and compound names with a native last part do have an -i in the dative:

Ég var að koma frá Noregi 'I just came from Norway'
Ég var að koma frá Finnlandi 'I just came from Finland'

The case of the neuter loanwords is used here for demonstration, as it is an instance of a systematic gap in the treatment of Icelandic inflection in the literature. More often than not, the lack is simply that of omission of information on individual words, if indeed it can be called omission, as the intention in the textbooks was probably never that of complete coverage. The problem is that the nature of a survey, which is at the core of the grammatical descriptions, is forgotten, and the rules intended for description are taken as absolutes.

The difference between the approaches to inflection between the textbooks and BÍN is as follows:

- The generalized rules in grammars and textbooks do not give enough information for individual paradigms. In BÍN, every single inflectional form has to appear.

- The scope of the textbooks is very narrow; they share most of the few examples they show. The vocabulary in BÍN is extensive, running into hundreds of thousands of headwords.
- There is a strong historical bias in the textbooks, giving obsolete and archaic forms their space, to the exclusion of newer words or variants, especially ones found to be unacceptable to purists. The aim in BÍN is to reflect actual use in the modern language, including loanwords and slang, irrespective of acceptability. Obsolete and arhaic forms are only included if they commonly appear in Modern Icelandic.

As the data and research on individual words is surprisingly scarce in the literature, the amount of research required for BÍN is considerable. In an online project this poses less of a problem than previously, although the users may have to get used to the fact that the description of the inflectional system is not something to be determined once and for all.

8. Conclusion

The work on linking ISLEX and BÍN is still at an experimental stage, but the hyperlinks are in place. ISLEX is expected to serve both native speakers of Icelandic and others needing an Icelandic-Scandinavian dictionary, whereas the twofold purpose of BÍN was centered primarily on language technology, and then, secondarily, on native speakers of Icelandic. BÍN is evolving into a multipurpose database, and as such it is difficult to serve all needs at once, the LT groups, native users, researchers, and students, at home and abroad. There is only one version of the database; it has to be as inclusive as it can be.

For BÍN to be as useful as possible to the users of ISLEX, the user interface needs to be translated to the languages used there. The website should also contain a manual and a thorough introduction to the methods used in the production of BÍN. These are minimum requirements, and they should not be to difficult to reach. The fact that non-native speakers might be better off with less information on variants, leaving out obscure or very specific instances, is harder to handle. In the original version of BÍN, variants were not linearly ordered. Now, an attempt has been made to order variants according to 'acceptability', and the notes accompanying the paradigms are intended to help the users cope. These notes, however, are probably of limited use to learners of Icelandic, as they assume native or near-native competence.

The linking of ISLEX and BÍN is therefore a break with traditions in Icelandic lexicography and inflectional description in two ways. Firstly, it gives the

users of ISLEX access to all the known facts on the inflection of individual words, perhaps sometimes to a point of superfluity, as in the instance of proliferating and/or obscure variants. Secondly, the attempt at describing the inflectional system 'bottom up' should lead to the discovery of the actual facts of language use, as in the case of the elusive dative ending in the multisyllable loanwords discussed above. That would at least give an alternative to the rule systems of the textbooks, which is sometimes based on partial data. As access to available data grows, with better tools of analysis, access to larger text collections and tagged corpora, etc., the picture should become clearer. BÍN is only the first step.

References

Árnason, Mörður (red.), 2007*:* Íslensk orðabók [Dictionary of Icelandic], 4. útg., Edda útgáfa hf., Reykjavík.

Jónsdóttir, Halldóra & Þórdís Úlfarsdóttir, 2012: ISLEX – en flersproget nordisk ordbog. I: Nordiska studier i lexikografi 11. Lund. (Skrifter utgivna av Nordiska föreningen för lexikografi 12.)

Kvaran, Guðrún, 2005: Orð. Handbók um beygingar- og orðmyndunarfræði. Íslensk tunga, II. bindi. Almenna bókafélagið, Reykjavík.

Magnússon, Friðrik, 1984: Ein lítil beygingarending. Mímir, Blað stúdenta í íslenskum fræðum 32:33–43.

Rögnvaldsson, E., H. Loftsson, K. Bjarnadóttir, S. Helgadóttir, A. B. Nikulásdóttir, M. Whelpton and A. K. Ingason, 2009: Icelandic Language Resources and Technology: Status and Prospects. Rickard Domeij, Kimmo Koskenniemi, Steven Krauwer, Bente Maegaard, Eiríkur Rögnvaldsson and Koenraad de Smedt (eds.): Proceedings of the NODALIDA 2009 workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources. S. 27–32. Northern European Association for Language Technology (NEALT), Tartu University Library.

BÍN = Beygingarlýsing íslensks nútímamáls. Kristín Bjarnadóttir (red). The Árni Magnússon Institute for Icelandic Studies 2002-, Reykjavík. http://bin.arnastofnun.is/

ISLEX. Halldóra Jónsdóttir, Þórdís Úlfarsdóttir (red.) The Árni Magnússon Institute for Icelandic Studies, Reykjavík. http://islex.hi.is; islex.dk; islex.no; islex.se.

Ritmálssafn Orðabókar Háskólans (The Written Language Archive): The Árni Magnússon Institute for Icelandic Studies, Reykjavík. http://arnastofnun.is/page/arnastofnun_gagnasafn_ritmal

Kristín Bjarnadóttir
The Árni Magnússon Institute for Icelandic Studies.
kristinb@hi.is