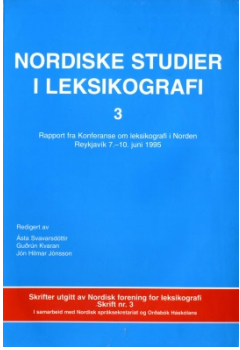


NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	Lexicalization and the Selection of Compounds for a Bilingual Icelandic Dictionary Base	
Forfatter:	Kristín Bjarnadóttir	
Kilde:	Nordiske Studier i Leksikografi 3, 1995, s. 255-263 Rapport fra Konferanse om leksikografi i Norden, Reykjavík 7.-10. juni 1995	
URL:	http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive	

© Nordisk forening for leksikografi

Betingelser for brug af denne artikel

Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Søgbarhed

Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for 'optical character recognition' og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Kristín Bjarnadóttir

Lexicalization and the Selection of Compounds for a Bilingual Icelandic Dictionary Base

Artikkelen handler om de problemer som hefter ved lemmatisering av sammensetninger i en tospråklig ordbok der islandsk er kildepråket. Fordi enkelte ord kan vise varierende ordformer som førsteledd i sammensetninger, vil lemmaseleksjonen ikke utelukkende gjenspeile semantisk leksikalisering. Det må også tas hensyn til at leksikaliseringen i mange tilfeller er begrenset til en bestemt formvariant. Dette forholdet kompliseres ytterligere ved at sammensetninger som viser et produktivt ordlagingsmønster, kan inneholde polyseme ordledd, eller ved at ordleddene står i en flertydig relasjon til hverandre.

1 Introduction

The subject of this paper is the selection of vocabulary for a bilingual Icelandic dictionary base, in particular the selection of compounds. As expected, lexicalization proves to be the major factor in the choice of compounds but, due to the complexities of Icelandic word formation, I will maintain that lexicalization can be said to apply to formal aspects as well as to the more widely discussed semantic ones.

My interpretation of lexicalization, then, is a rather extensive one: In the following pages I will be assuming that lexicalization simply refers to all features which can not be inferred from the sum of the parts of any process of word formation, including ambiguity of form.

2 The Project

The project I am describing is the creation of an Icelandic base for a bilingual (i.e. Icelandic-Scandinavian) dictionary, being worked on at present at The Institute of Lexicography. The project is a joint venture of The Institute and Nordisk språksekretariat, and it is jointly funded by The Institute and Nordisk kulturfond. The intended size of the base is approximately 50 000–60 000 headwords. The work is to proceed in stages; the first stage, presently in progress, is the macrostructure; that is the word list itself, with information on inflections, word formation, idioms and other phrasal entries, etc.

The idea is to complete the work on the macrostructure as far as possible before turning to the description of individual headwords or the microstructure of the dictionary base.

At the outset a pilot word list containing just under 180 000 words was assembled from selected sources, and the observations made in this paper are drawn from the practical problems encountered while trying to reduce that number to the intended 50 000–60 000 words.

2.1 The Primary Sources

The main sources of the pilot word list are the archives of The Institute of Lexicography, a recent frequency dictionary of modern Icelandic (Jörgen Pind et al. 1991), compiled and published by The Institute, and the combinatorial dictionary *Orðastaður* described by the author Jón Hilmar Jónsson (1994) in his paper in this issue. As stated above, the basic word list arrived at from these three sources contains approximately 180 000 words.

As an indication of the scope of these sources the largest of the Institute's archives, *The Written Language Archive* (WLA), contains well over 600 000 headwords, the frequency dictionary contains just over 31 000 lemmas, and Jón Hilmar's book contains 11 000 headwords and over 100 000 compounds. These three sources are computerized so that producing the pilot list was relatively simple.

It should be noted that from the outset it was clear that some fields were underrepresented in the original data, which was only to be expected, given the nature of the sources. The historical aspect has always been in the foreground at the Institute, which in fact was founded just over fifty years ago for the purpose of producing a historical dictionary of Icelandic from the beginning of the age of printing in Iceland, i.e., 1540, to the present day. The WLA was created by excerpting for that purpose, and not intended as a base for a dictionary of the modern language (cf. Ásta Svavarsdóttir/Jón Hilmar Jónsson/Kristín Bjarnadóttir 1991). The combinatorial dictionary largely shows productive word formation, and the frequency dictionary, despite its rather hefty size, is based on a limited corpus of texts. The preliminary list of 180 000 words thus had to be reduced even further than down to 50 000–60 000, to leave room for new material. One of the sources of new material is the Institute's text corpus, and some excerpting in the traditional manner will also be done to supplement the original data. The bulk of the material will, however, come from the three sources mentioned above.

2.2 The Sorting of the Data

The sorting of the words in the original list was started in March 1994, and by now the whole alphabet has been sorted. Out of the 180 000 words in the original list, there are just over 138 000 compounds, and just over 35 000 derived or simple forms, i.e., non-compounds. Just over 4 500 "words" or items could not be analysed in this manner; these included abbreviations, some phrasal constructions, etc., as well as some items that can really only be classified as accidentals!

The sorting codes are kept as simple as possible, and the words are marked *included*, *omitted* and *under consideration*. Please note that *omitted* does not imply that the word is deleted from the lists, just that it does not get the status of a prospective headword.

Table 1 shows the proportions of base words, derived words, and compounds in the original word list, and the percentages in the sorted data for each type of word formation category, as they now stand in the lists:

	Base words		Derived words		Compounds	
To be included	8 276	44%	5 675	35%	27 247	20%
Under consideration	2 096	11%	2 290	14%	12 483	9%
To be omitted	8 555	45%	8 226	51%	98 495	71%
	18 927		16 191		138 225	

Table 1: *Proportions in the Pilot List*

The division between words to be included and those to be left out is not, of course, as clear-cut as these figures might imply. The compounds are in fact sorted into four categories, two of which contain candidates for our targeted 60 000 words. These categories were arrived at after quite some experimenting. Very simply put, the first category contains regularly formed compounds that are part of the core vocabulary, and common compounds in which meaning and form are not predictable, i.e., lexicalized words. The third category contains compounds which in fact do require explanation or listing, without being candidates for a base of the size intended here, as they are more peripheral in some manner, very often rare, specialized, old, dialectal, etc. The second category falls between the two, and the treatment of these words will be a matter of editorial policy in the future. The fourth category contains the words that do not really need explanations, being the products of fully productive and unambiguous word formation rules. Examples of this very simple outline of the classifications are given in (1), i.e., examples of compounds with *blóð* 'blood' as the first constituent.¹ All the examples seem to have the same morphological construction, and formally they are perfectly regular. The senses range from being fully lexicalized, as in (1a) *blóð-berg* 'thyme' (*timian*), to being fully productive as in (1f) *blóð-lykt* 'the smell of blood'.² The figures for each of the categories of compounds as they now stand in our files are given in parenthesis preceding the examples.

(1) *Examples of compound classification***Lexicalized compounds:**

INCLUDED (27 247):

- a *blóð-berg* 'Thymus arcticus'
- b *blóð-steinn* 'hematite'
- c *blóð-suga* 'blood sucker, vampire'

UNDER CONSIDERATION (12 483):

- d *blóð-arfi* 'Polygonum aviculare'
- e *blóð-miga* 'hematuria'
- f *blóð-rót* 'Potentilla erecta'

¹Please note that for reasons of clarity a hyphen has been inserted between the component parts of compounds in all examples, although this is not in accordance with Icelandic spelling. Please note also that the glosses are intended as rough guides to the meaning of the words, as it is sometimes impossible to give the exact meaning without giving lengthy definitions. Some words proved to be completely beyond translation except in context, and these are left without glosses.

²The ordinary user is not going to know that *blóð* in the first word is probably not the same as in the second. Cf. Ásgeir Blöndal Magnússon 1989.

OMITTED (13 529):

- g *blóð-björg* rarer name of *blóðberg*
- h *blóð-lýsa* 'leukemia'
- i *blóð-tala* 'red blood cell'

Productive compounds:

OMITTED (84 966):

- j *blóð-blettur* 'blood stain'
- k *blóð-bragð* 'taste of blood'
- l *blóð-lykt* 'smell of blood'

Independently of this very simple classification, the words are also sorted according to the field in which they are used, e.g., according to academic subject (*botany, zoology, medicine, etc.*), craft or occupation (*carpentry, sewing, agriculture, etc.*), and various other fields of diverse kinds (*literature, music, art, theatre; cars, traffic, flying; food, toys, etc.*) There are even fields for *sheep, cows* and *Icelandic national costume!* These fields will be used to make the classification more cohesive and consequent, thus giving a thematic key to the vocabulary.

2.3 The Criteria for the Sorting

The examples *blóð-berg* and *blóð-lykt* represent the two ends of the spectrum. One is fully lexicalized and the other is formed by a fully productive word formation rule. Lexicalizations of this kind are included in dictionaries or not, solely on the grounds of currency, i.e., whether the word is central enough in the vocabulary. As regards the productive part of compounding, the lexicographer is faced with the problem of deciding how far the prospective user can be expected to handle productivity, and how his needs can best be met.

The problem with sorting of the kind described here is of course that there are very many features that have to be taken into account when trying to determine what the prospective user may need. The problem is even greater when the prospective user is as remote as in this project. We are making a base for a bilingual dictionary which has to be useable for more than one set of languages. It is clear that the criteria are bound to be quite different according to the purpose of the end result, and we found it very important to try to construct the base in such a way that it could be used in as flexible a manner as possible. We also did not want the classification to be too complicated, for obvious reasons of time and money. We are therefore using a simple classification system to cope with very complex matters.

3 A Few Types of Compounds

Fully lexicalized words, like the examples I have been using, seem to have nothing whatsoever in common with their component parts. Fully productive compounds, on the other hand, seem to be nothing but the sum of their parts. Life would be relatively easy if that was all there was to it, but that is of course not the case.

3.1 Ambiguous Elements

Word formation rules are not unambiguous, and the ambiguity can appear both in the components themselves and the relation between them. Some regularly formed words are therefore not fully transparent.

In the examples in (2) below one of the component parts, *saumur*, can mean 'stitch', as in needlework, for example in *flat-saumur* 'flat stitch' or 'satin stitch', but *saumur* is also used in carpentry for some types of nails, such as *þak-saumur* and *pappa-saumur*.

- (2) a *flat-saumur* 'flat stitch', 'satin stitch'
 b *þak-saumur* 'roofing nail'
 c *pappa-saumur* 'nail used to fasten tar paper'

All homonyms are therefore a source of confusion when it comes to compounds.

The terminology in very many fields in Icelandic abounds with such words, as language policy dictates the preference of Icelandic neologisms over the use of loan words. "Ordinary" words are therefore very often used with specialized meanings, and the elements can thus very easily be semantically ambiguous.

The combining forms can be ambiguous as well, as shown in (3) and (4) below. The words *önd* 'duck' and *andi* 'spirit' can both have the combining form *anda-*. The accepted meaning of the compound in (4) *anda-læknir*, is 'spiritual healer', but I have seen it used in a police report to refer to 'a veterinarian specializing in treating ducks'.³

- (3) a *önd* fem. 'duck' Combining form: *anda-*
 b *andi* masc. 'spirit' Combining form: *anda-*
- (4) a *anda-læknir* 'spiritual healer'
 b *?anda-læknir* 'veterinarian specializing in ducks'

Although such examples can be humorous, they can also be the cause of very real and even quite serious misunderstandings.

3.2 Ambiguous relations of elements

An ambiguity which is perhaps not likely to cause confusion of a similar magnitude as the one mentioned above, is the ambivalence in the relation between the elements in a compound. This is demonstrated in the examples in (5) below.

- (5) a *lauk-baka* 'onion pie'
 b *sælkerabaka* 'gourmet pie'

The difference between 'onion pie' and 'gourmet pie' is obvious, but both words are regularly formed. Examples such as these are not likely to cause any difficulties for anyone, but not all cases are as obvious as these. In the sorting of our material we have found that it is necessary to be aware of these ambiguities in order not to overlook some less than obvious meanings.

³The story was that a sick duck was reported to the police. It got better on its own before the police found an *anda-læknir* for it.

3.3 The Choice of Lexical Items

The last but one of the types of compounds I will mention are words that are perfectly regularly formed, with perfectly transparent meaning, but they are lexicalized in the sense that the choice of components is not free. These are very often a source of mistakes for foreigners. To name an example (6), the words *verslun* and *búð* are synonyms for the word 'shop', although the first one is more formal than the second. Both *verslun* and *búð* can be used for places that sell books (as in (6a)) or flowers or cosmetics. But for an 'ice cream parlour' only *-búð* is acceptable, *ís-búð*.

- | | | | | |
|-----|---|-----------------|------------------------|---------------------|
| (6) | a | <i>bóka-búð</i> | <i>bóka-verslun</i> | 'bookstore' |
| | b | <i>fisk-búð</i> | ?? <i>fisk-verslun</i> | 'fishmonger's' |
| | c | <i>ís-búð</i> | * <i>ís-verslun</i> | 'ice cream parlour' |

The reaction of native speakers to mistakes made by foreigners in compounds such as these is usually "But you can't say that, actually, the Icelandic is ...". When pressed for an explanation, the only answer usually is: "Það er bara svona!" "That's just the way it is!"

3.4 Differences in Form

At the beginning of this paper I claimed that some formal aspects, such as differences in combining forms in Icelandic, have to be considered lexicalizations.

The most striking sets of combining forms do not in fact occur in compounds but in derivations (cf. (7)).⁴ As the glosses show, the meaning of each of the three words formed with the word *maður* 'man' and the affix *-legur* '-ly' is quite distinct from the other two:

- | | | | | |
|-----|---|---------------------|--------------------|-------------------------|
| (7) | a | <i>mann-legur</i> | <i>manns-legur</i> | <i>manna-legur</i> |
| | | 'human' | 'manly' | 'precocious, conceited' |
| | b | ? <i>barn-legur</i> | <i>barns-legur</i> | <i>barna-legur</i> |
| | | | 'childlike' | 'naive, childish' |
| | | | 'earnest, sincere' | |

Examples like these, where different combining forms are distinctive as to meaning, are very rare. More commonly only some of the possible combining forms are acceptable, even when the word formation is fully productive in all other respects. In order to explain this point a little excursion into Icelandic word formation is needed.

4 A Brief Word on Icelandic Word Formation

Contrary to what many recent publications on word formation maintain,⁵ Icelandic seems to contain word internal inflectional endings. A noun appearing as the first part of a compound can thus have the form of a stem or an oblique case, usually the genitive, or, more rarely, dative. Link phonemes do also occur. Adjectives, as first parts of compounds, contain inflectional endings as well, some of which are inflected for case inside the compounds. The

⁴Although the status of the affix *-legur* is a bit dubious on some formal grounds, at least it fulfills the requirement of being a bound form.

⁵I am mainly referring to discussions on Level Ordering.

verbal morphology in compounds is quite varied too. As noun+noun compounds are the most common and the most varied I will use these to demonstrate my point.

Examples (8) to (13) show some compounds with the nouns *bók* 'book' and *barn* 'child' as the first component part. Both these words commonly appear in compounds in the three forms shown in the examples. The choice of form for the first element in the compound seems to be largely arbitrary, as seen in (8a) and (8b). There is no semantic reason for the word 'book' to be in the plural in *bóka-búð* but not in *bóka-sala*, and the reason cannot be phonotactic either.

Noun+Noun Compounds:

	Stem:	Genitive plural:	
(8) a	<i>bók-sala</i>	<i>bóka-búð</i>	'bookstore'
	* <i>bók-búð</i>	* <i>bóka-sala</i>	
b	<i>bók-hlaða</i>	<i>bóka-safn</i>	'library'
	* <i>bók-safn</i>	* <i>bóka-hlaða</i>	
	Stem:	Genitive sg.:	Genitive pl.:
(9)	<i>bók-merki</i>	<i>bókar-merki</i>	<i>bóka-merki</i> 'book mark'

Words like the ones in (9) *bók-merki*, *bókar-merki*, *bóka-merki* 'book-mark' where more than one form is acceptable to convey the same meaning are common, as shown by the fact that The Institute gets a great number of phone calls from people asking about "the correct form".⁶ Part of the problem is perhaps that the difference can be quite difficult to hear, as in *bókar-merki* and *bóka-merki*. People have to be quite articulate for that distinction to be heard.

In examples (10) to (13) the reasons for the differences in forms cannot be semantic either. The difference between (13a) and (13b) is that *barns-meðlag* is 'child support' paid by a parent, whereas *barna-lífeyrir* is a part of the social security system. Both can be used to apply to one or more children.

(10) a	Stem:	<i>barn-fóstra</i>	'children's nurse'
b	Gen.sg.:	<i>barns-faðir</i>	'father of a child'
c	Gen.pl.:	<i>barna-pía</i>	'babysitter'
(11) a	Stem:	<i>barn-æska</i>	'childhood'
b	Gen.sg.:	<i>barns-aldur</i>	'childhood'
(12) a	Gen.sg.:	<i>barns-vagga</i>	'crib'
b	Gen.pl.:	<i>barna-rúm</i>	'baby's bed'
(13) a	Gen.sg.:	<i>barns-meðlag</i>	'child suport'
b	Gen.pl.:	<i>barna-lífeyrir</i>	'child support'

In an ideal system, the distribution of genitive singular and plural would obey the same rules as in syntax, where the feature *number* would be used consequently. Icelandic indeed has such words, as shown by the words for the son(s) and daughter(s) of farmer(s), brother(s) and king(s) below:

⁶A similar observation on Swedish is made in Malmgren 1994.

- | | | | | |
|------|---|-----------------------------|---|-----------------------------|
| (14) | 1 | $N_{gen.sing.} + N_{sing.}$ | 2 | $N_{gen.sing.} + N_{plur.}$ |
| | a | <i>bónda-sonur</i> | a | <i>bónda-synir</i> |
| | b | <i>bróður-sonur</i> | b | <i>bróður-synir</i> |
| | c | <i>konungs-dóttir</i> | c | <i>konungs-dætur</i> |
| | 3 | $N_{gen.plur.} + N_{sing.}$ | 4 | $N_{gen.plur.} + N_{plur.}$ |
| | a | ? <i>bænda-sonur</i> | a | <i>bænda-synir</i> |
| | b | * <i>bræðra-sonur</i> | b | <i>bræðra-synir</i> |
| | c | ? <i>konunga-dóttir</i> | c | <i>konunga-dætur</i> |

The words marked with a question mark, (14-3a) *bændasonur* ‘the son of farmers’ and (14-3c) *konungadóttir* ‘the daughter of kings’, are not very probable for semantic reasons, but not completely impossible. The first one is slightly better than the second one, as there are female farmers these days. The problem with the second one is that the word for *king* in Icelandic is not only a masculine noun but is used of males only, unlike the words *forseti* ‘president’ and *borgarstjóri* ‘mayor’ that are masculine nouns but readily used of women. The word marked with a star, (14-3b) *bræðra-sonur* ‘the son of brothers’, is obviously impossible for semantic reasons!

Words such as these are rare, whereas examples such as the ones in (8) to (13) are much more common. Usually the difference in form is just exactly that, an arbitrary difference. It is a matter of accepted forms and those that just sound wrong. Why they do is not a question that is easily answered.

4.1 The Distribution of Forms

The search for phonotactic, semantic, or formal explanations for the difference in combining forms only turns up isolated explanations. Some inflectional classes only seem to appear in certain morphological constructions. Feminine nouns ending in *-a* in the nominative singular only use the genitive singular as a combining form, even in nouns such as the one in (15a) *peru-tré* ‘pear-tree’ when the genitive plural would perhaps be expected semantically, as seen in (15b) *epla-tré* ‘apple tree’, where the apples appear in the plural.

- | | | | | |
|------|---|--------------------|----------------------|--------------|
| (15) | a | <i>peru-tré</i> | * <i>per(n)a-tré</i> | ‘pear tree’ |
| | | [fem.sing.] | [fem.plur.] | |
| | b | * <i>eplis-tré</i> | <i>epla-tré</i> | ‘apple tree’ |
| | | [neut.sing.] | [neut.plur.] | |

Another indication of a rule is that in multiple compounding the tendency seems to be to use the genitive form when the first part of the multiple compound is a compound itself (cf. Baldur Jónsson 1984, and Eiríkur Rögnvaldsson 1986), as demonstrated in (16a):

Multiple compounds as first component parts:

- | | | | | |
|------|---|--------------------|---|-------------------------------------|
| (16) | a | [[[N][N]]] | b | [[[N][N]] _{gen.sing.} [N]] |
| | | <i>borð-plata</i> | | <i>skrifborðs-plata</i> |
| | | ‘table top’ | | ‘writing table top’ |
| | | | | i.e. ‘desk top’ |
| (17) | a | <i>verð-hækkun</i> | b | <i>olíuverðs-hækkun</i> |
| | | ‘price rise’ | | ‘oil-price rise’ |

Even though there are some rules on a par with the two just mentioned, the problem is that very many factors play a role, and in the end the conclusion always seems to be: That's just the way this word is. That, I maintain, is a clear indication of lexicalization.

5 Conclusion

This very brief outline of some of the problems posed by the complexity of the formal aspects of Icelandic compounding is hopefully sufficient to show that the subject deserves careful attention.

As we are not producing a fully fledged dictionary, but rather a base from which dictionaries are to be produced in the future, our solution is to list all combining forms under the words they are derived from. The combining forms themselves are then graded for inclusion or omission in the base, in the same way as the compounds, or indeed all other words from the pilot word lists. All compounds are then placed under the proper combining form. The lists are therefore not in alphabetical order, but ordered morphologically, or rather indexed according to the morphology. This means that the word in (3a) *önd* 'duck' is not separated by the whole alphabet from its combining form *anda-*.

I should stress again that productively formed compounds (and derivations) are very much a part of the dictionary base. They are not deleted from the lists, just marked to show that they are not expected to be given the status of headwords.

It will be a matter of editorial policy in the future just how the base is developed and used. The relative weight of the factors I have mentioned will of course vary according to the aims of the editors, together with the question of how active or passive the prospective dictionary is supposed to be. For now we just have to make sure that the base is flexible enough to serve as many needs as possible.

Bibliography

- Ásgeir Blöndal Magnússon 1989: *Íslensk orðsifjabók*. Reykjavík: Orðabók Háskólans.
- Ásta Svavarsdóttir/Jón Hilmar Jónsson/Kristín Bjarnadóttir 1992: Fra seddelsamling til database: Leksikografisk analyse af islandske verber. In: Fjeld, R. V. [ed.]. *Nordiske studier i leksikografi*, 390–402.
- Baldur Jónsson 1984: Samsett nafnorð með samsetta liði. Fáeinar athugasemdir. In: *Festskrift til Einar Lundey*. 3. október 1984, 158–174.
- Eiríkur Rögnvaldsson 1986: *Íslensk orðhlutafraeði. Kennslukver handa nemendum á háskólastigi*. Reykjavík: Málvísindastofnun Háskóla Íslands.
- Jón Hilmar Jónsson 1994: *Orðastaður. Orðabók um íslenska málnotkun*. Reykjavík: Mál og menning.
- Jón Hilmar Jónsson 1995: Nøkler til ordforrádet: Om lemmafunksjon, struktur og informasjonstyper i en ny kombinatorisk ordbok over islandsk. [This volume, 245–254].
- Jörgen Pind [ed.]/Friðrik Magnússon/Stefán Briem 1991: *Íslensk orðtíðnibók*. Reykjavík: Orðabók Háskólans.
- Malmgren, Sven Göran 1994: Sammensättningsmorfologi och lexikografi. In: Anna Garde/Pia Jarvad (red.): *Nordiske studier i leksikografi II. Rapport fra Konference om Leksikografi i Norden 11.–14. maj 1993*, 179–184.