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

Kristín Bjarnadóttir
The Árni Magnússon Institute for Icelandic Studies

The 11th NFL Conference in Lund,
26 May 2011

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

The linking of two projects at The Árni Magnússon Institute for Icelandic Studies

ÍSLEX

- The Icelandic /Danish, Norwegian and Swedish electronic dictionary discussed by Halldóra Jónsdóttir and Þórdís Úlfarsdóttir in their talk on Friday, May 27th
- ÍSLEX contains 50,000 headwords

BÍN - The Database of Modern Icelandic Inflection [Beygingarlýsing íslensks nútímamáls]

- 270,000 paradigms
- 5.8 million inflectional forms
- Work on BÍN started in 2002

Breaking away from tradition . . .

- By giving much more information on inflection than possible in Icelandic dictionaries until now.

- By using a ‘bottom up’ approach to describe the Icelandic inflectional system, instead of the ‘top-down’ approach traditionally used in grammatical description.

  - Textbooks are intended to give a survey of the system as a whole.
  - Paradigms for individual words have to show the inflection, “as is”, i.e. actual usage.

About BÍN

- The **Database of Modern Icelandic Inflection** was originally intended for NLP use, both for analysis and production.

- The aim was to show "all" and only existing word forms from the modern language in each paradigm.

- The paradigms are accessible on the ÁM Institute's website. Visits in the last 12 months: **398,768** (326,559 in Iceland)
  Unique visitors: **82,752**

- The users of the website are mostly Icelandic, and they expect the database to show "accepted usage".

- The NLP needs and the expectations of the public are not necessarily the same.

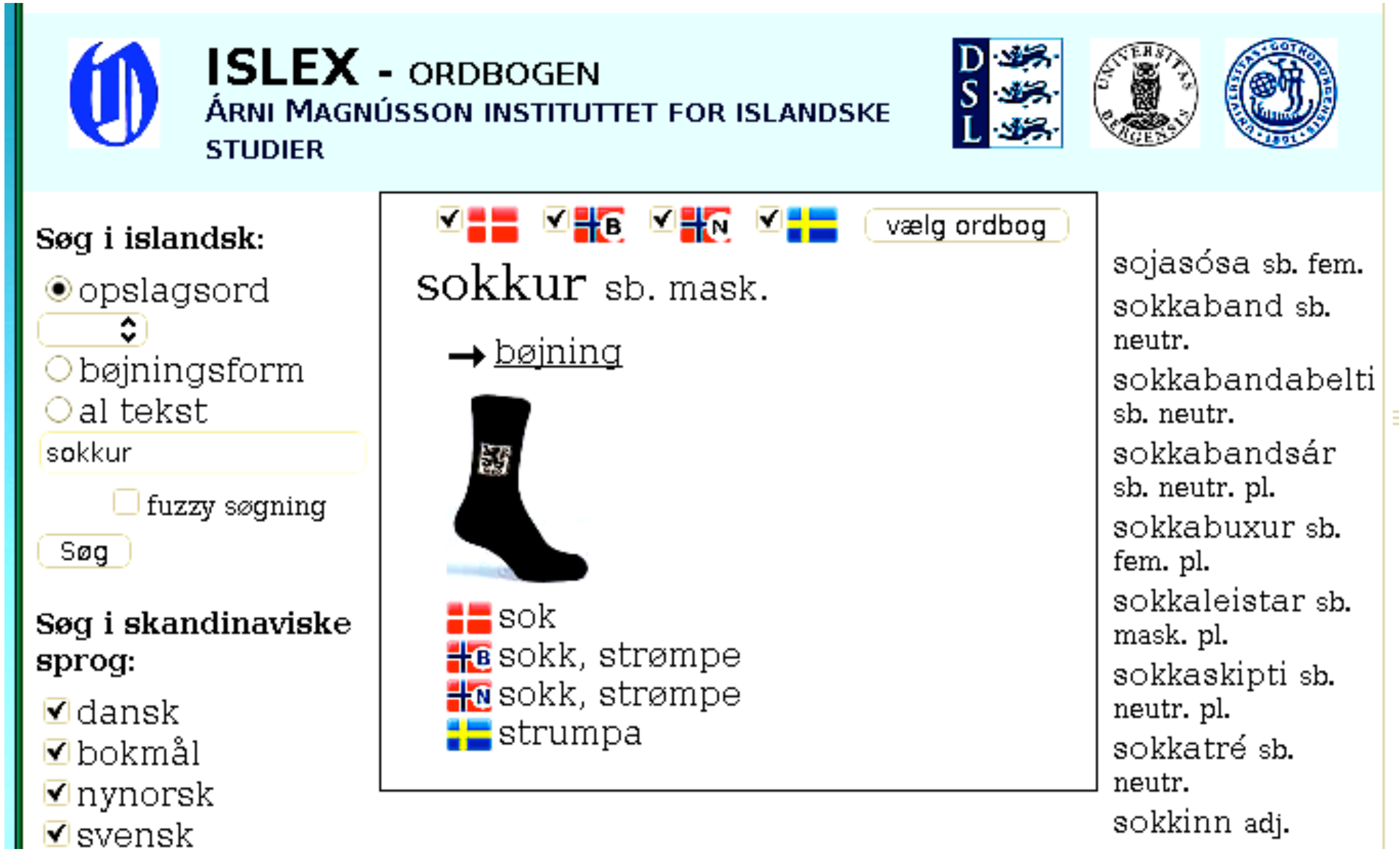
Inflection in Icelandic is quite complex

The number of inflectional forms in the paradigms

- **Nouns: 16** forms (4 cases, singular and plural, +/-definite)
- **Adjectives: 120** forms (3 genders, 4 cases, singular and plural, +/-definite; positive, comparative, superlative)
- **Verbs: 107** forms, (indicative/subjunctive, present/past, person, number, etc.)

The figures are for a full paradigm, **excluding variants**.

There are over 550 inflectional classes in the database.
These are a tool for maintaining the database.
They are not accessible on the website.



ISLEX - ORDBOGEN

ÁRNI MAGNÚSSON INSTITUTTET FOR ISLANDSKE STUDIER

Søg i islandsk:

● opslagsord

○ bøjningsform

○ al tekst

sokkur

□ fuzzy søgning

Søg

Søg i skandinaviske sprog:

☑ dansk

☑ bokmål

☑ nynorsk

☑ svensk

☑ ☑ B ☑ N ☑ vælg ordbog

sokkur sb. mask.

→ bøjning

	sok
B	sokk, strømpe
N	sokk, strømpe
	strumpa

sojasósa sb. fem.

sokkaband sb. neutr.

sokkabandabelti sb. neutr.

sokkabandsár sb. neutr. pl.

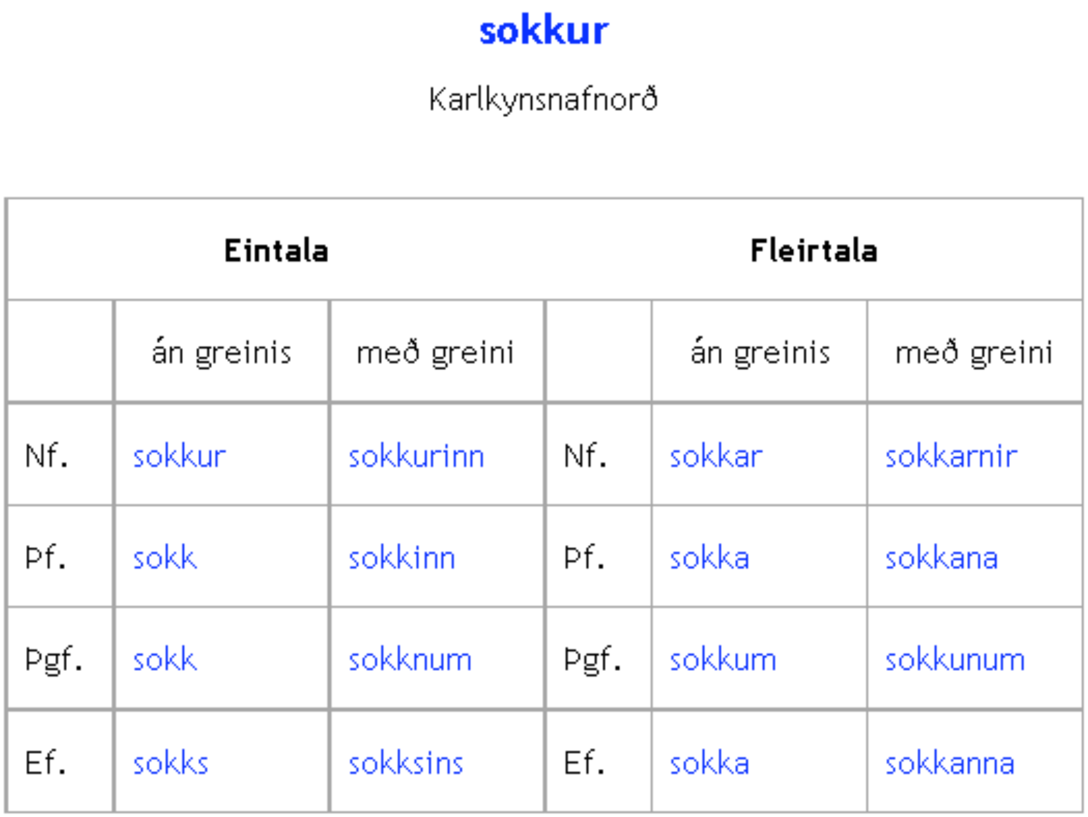
sokkabuxur sb. fem. pl.

sokkaleistar sb. mask. pl.

sokkaskipti sb. neutr. pl.

sokkatré sb. neutr.

sokkinn adj.



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

The linking of ISLEX and BÍN

All the relevant headwords in ÍSLEX are simply linked to their paradigms in the BÍN Database, using hyperlinks with the id of each paradigm.

NB:
BÍN is also used in searching ISLEX:

köttur kött ketti kattar kettir köttum katta kötturinn köttinn kettinum kattarins kettirnir kettina köttunum kattanna
→ köttur ‘cat’

Breaking away from tradition . . . I.

By giving much **more information** on inflection than possible in Icelandic dictionaries until now.
The tradition is to give “principle parts” only:

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**

Why?
The principle parts are not sufficient to predict the whole paradigm.

The unpredictable parts of ‘sokkur’, given the principle parts “-ar, -ir”

		Singular, indef.		
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’

Dat.sg.definite:
sokknum, flokknum, lokknum, kokkinum/kokknum

“Allt í kleinu hjá *kokknum* ...”

The variants in BÍN ... as in lokk/lokki

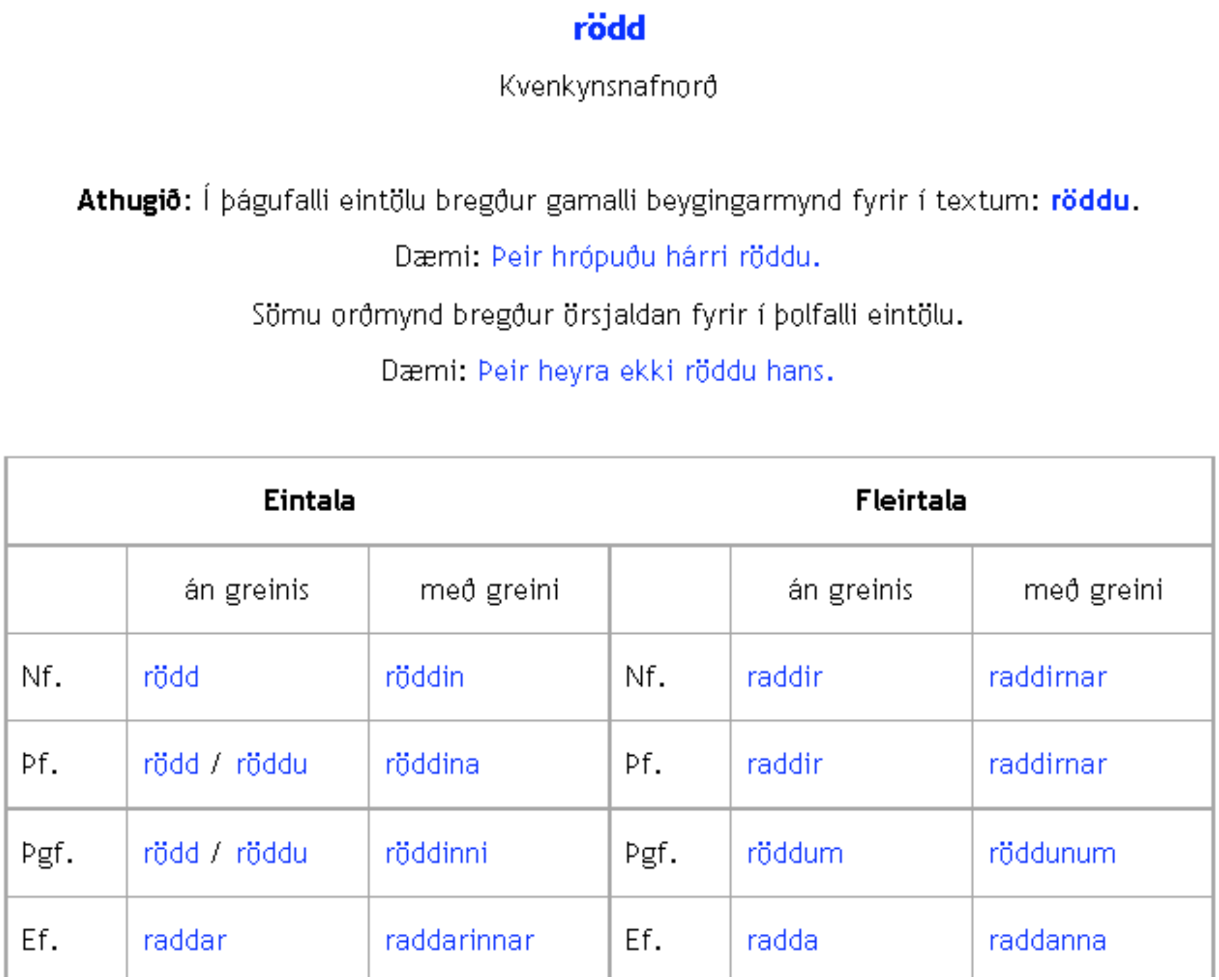
- The actual paradigms list all variants, without differentiation for usage, style, etc. There is no change in font, colour, etc.

- For NLP use, the database has to be as inclusive as possible. E.g.: If a word form is not in the database, it will not be included in the input for a search engine.

- Native speakers also want information on the variants, not just a list of them.

- Therefore, notes on choice of variants, usage, etc. are included at the top of the paradigm.

- The notes are intended to help the user pick the right variant, or at least to keep him from using highly restricted ones, cf. **rödd** ‘voice’ →



rödd

Kvenkynsnafnorð

Athugið: Í þágufalli eintölu bregður gamalli beygingarmynd fyrir í textum: **röddu**.

Dæmi: Þeir hrópuðu hárri röddu.

Sömu orðmynd bregður örsjaldan fyrir í þolfalli eintölu.

Dæmi: Þeir heyra ekki röddu hans.

Eintala			**Fleirtala**		
	án greinis	með greini		án greinis	með greini
Nf.	rödd	röddin	Nf.	raddir	raddirnar
Þf.	rödd / röddu	röddina	Þf.	raddir	raddirnar
Þgf.	rödd / röddu	röddinni	Þgf.	röddum	röddunum
Ef.	raddar	raddarinnar	Ef.	radda	raddanna

The note on ‘rödd’ in BÍN

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.’

Breaking away from tradition . . . II:

Traditional descriptions of the Icelandic inflectional system:

Surveys: The aim is to present the whole as clearly as possible, top-down.

- Classification according to **principle parts** only.
- A **limited number of inflectional classes** (with notes on exceptions).
- A **limited number of examples**, chosen to demonstrate the system as a whole.
- The emphasis on ‘good Icelandic words’.

There are gaps in the treatment of Icelandic inflection, as some aspects are not present in the literature.

Example of a gap in the grammar books:

The endings -i/-0 in the dative singular of neuter nouns

• The traditional literature: The ending **-i** is (almost) **universal** in the dative singular in neuter nouns. The **only** exceptions are **tré** ‘tree’, **fé** ‘money; livestock’, **hné/kné** ‘knee’.

This is in fact true for single syllable and affixed words from the inherited Icelandic vocabulary.

• Loanwords, especially multisyllable words, very often do not have an ending in the dative singular (-0). These words are omitted in most grammars.

Some of these loanwords are by no means new: **fenikel, kanel, vítríól, stúdíum, examen, fíaskó** and **ópíum** are attested in the Institute’s archives at the turn of the 18th C, i.e. ‘before Rasmus Christian Rask’.

The 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'

- Cf. older names of countries which do have **-i**:
  Ég var að koma frá Noregi 'I just came from Norway'
  ... and compounds with a native last part:
  Ég var að koma frá Finnlandi 'I just came from Finland'

‘Top-down’ to ‘bottom-up’

- The generalized rules in grammars and textbooks do not give enough information for individual paradigms.

- The scope of the textbooks is very narrow; they share most of the few examples they show.

- There is a strong historical bias, giving obsolete and archaic forms their space, to the exclusion of newer variants.

- The aim of BÍN is the production of individual paradigms, reflecting actual use in the modern language.

- The data and research on individual words is surprisingly scarce.

The data on “sokkur”

- **Generalizations** abound in discussions of inflections. People will stick to the rules in the grammar books when answering questions on usage, even though they may never actually use the inflectional forms they consider or mark as ‘correct’. The grammar books suggest **-i** for masculine nouns ending in **-ur**.

- The dative form **sokk** is attested in all the examples in the archives at the SÁ Institute, in the SÁ text collection (65 million words), and the available sources from Old Icelandic. The web supports this, although I did manage to find one example of “**sokki**” on the web (other than a horse named “Sokki”).

Typical example: **Ég fór úr blautum sokk/*sokki**
‘I took off a wet sock’

Future work on linking ISLEX and BÍN

- The work is still in an experimental stage.

- The user interface needs to be translated to the languages used in ISLEX.

- The website should contain a manual and a thorough introduction to the methods used in the production of BÍN.

- The opinions of the users, as evidenced in emails and comments, should be considered carefully.

BÍN is a multipurpose database, and as such it is difficult to serve all needs at once, the NLP groups, researchers, native users, and students, at home and abroad.

Using BÍN for non-native speakers

- The **abbreviations** in the paradigms are in Icelandic (but that is not too difficult to remedy)

- Non-native speakers might be better off with **less information** on variants, leaving out obscure or very specific variants.

  Example: The variant **kokknum** (‘chef’ dat.sg.+def.) could be left out, it is quite marginal.

- In order to make the material more accessible to website users, the variants are **ordered** according to ‘acceptability’. (In the original NLP version, variants were not ordered.)

There is only one version of the database; it has to be as inclusive as possible.

Thank you for your attention!

Kristín Bjarnadóttir
BÍN

kristinb@hi.is
http://bin.arnastofnun.is