Kristín Bjarnadóttir & Jón Friðrik Daðason
The Árni Magnússon Institute for Icelandic Studies/
University of Iceland

Standardization, prescription, description and theory: Icelandic inflection

25th Scandinavian Conference of Linguistics,
Workshop 2: Foundations of Language Standardization

Reykjavík 14th May 2013

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

The topic

The effect of extensive empirical data on the traditional descriptions of Icelandic inflection,

- using the dative singular **i/-0** variants in strong neuter nouns as a focus,
- the implications of such research on the standardization of Icelandic inflection.

The inspiration: Brown, Dunstan, & Roger Evans. 2012. Morphological complexity and unsupervised learning: Validating Russian inflectional classes using high frequency data.

We are not into machine learning in this project yet!

Outline

- The dative singular of strong neuter nouns
- The standard, the prescription: Grammar books
- The variants, the description: The Database of Modern Icelandic Inflection
- The project: Finding and analysing the data in corpora
- Results
- Implications for standardization
- Implications for grammatical theory?

The hypothesis: Prescriptive grammars should be based on research.

The dative singular of neuter nouns in Icelandic

The simplest part of Icelandic inflection: Neuter nouns
Strong inflection: **-i** in the dative singular:

barn ‘child’: dat.sg. **barni**
skáld ‘poet’: dat.sg. **skáldi**
ljóð ‘poem’: dat.sg. **ljóði**

The standard textbooks mention 3 exceptions to the rule:

fé ‘money; sheep’: dat.sg. **fé**
tré ‘tree’: dat.sg. **tré**
hné ‘knee’: dat.sg. **hné**

Work on the Database of Modern Icelandic Inflection (BÍN) has shown that the **-i** is not a universal dative ending in strong neuter nouns:

bíó ‘cinema’: dat.sg. **bíó/bíói**
helíum ‘helium’: dat.sg. **helíum**
. . .

The Classic Grammar: Valtýr Guðmundsson. 1922.

Intetkønsordene danner kun én Klasse, idet de alle har -s í G. Sg. og er uden Endelse i N. Pl. Men da deres Bøjning iøvrigt kan være ret forskellig, maa der opstilles ikke mindre end 8 Bøjningsmønstre. Som saadanne kan tjene: *borð* Bord, *land* Land, *kyn* Køn, Slægt, *trje* Træ, *hreiður* Rede, *Meðal* Middel, Medecin, *klæði* Klæde, *ríki* Rige:

Sg.	N.	borð	land	kyn	trje
	A.	borð	land	kyn	trje
	D.	**borð-i**	**land-i**	**kyn-i**	**trje**
	G.	borð-s	land-s	kyn-s	trje-s

Pl. [. . .]

Sg.	N.	hreiður	meðal	klæði	ríki
	A.	hreiður	meðal	klæði	ríki
	D.	**hreiðr-i**	**meðal-i**	**klæði**	**ríki**
	G.	hreiður-s	meðal-s	klæði-s	ríki-s

Pl. [. . .]

(Colours added)

Characteristics of the grammar books

- Classification into inflectional classes based on principle parts, i.e., nom.sg., gen.sg. and nom.pl. for nouns.
- Small vocabulary, with (some) exceptions listed. Generally the same examples throughout the literature.
- The aim is a clear picture with (more or less) universal rules, i.e., a top-down survey of the inflectional system.
- Due to the influence of language purism: A strong emphasis on the core vocabulary from Old Icelandic.

The grammar books are used as a basis for standardization, e.g.,:

3. Þágufall eintölu sterkrar beygingar endar alltaf á *-i*.
'**The dative singular strong inflection** [of n.neut.] **always ends in –i**.'

(Guðrún Kvaran. 2005. *Orð*. Handbók um beygingar- og orðmyndunarfræði. [Words. Handbook of inflection and word formation.] p:244.)

(The exceptions *tré* 'tree', *hné* 'knee' and *fé* 'money; sheep' are shown on pages 242, 246.)

Counterexamples

Jón var að koma frá **Pakistan/*Pakistani**<dat>
’Jón has just come from Pakistan’

Cf. Jón var að koma frá **Englandi/*England**<dat>
Jón has just come from England’

Þorskur með **saffran/saffrani**<dat> fyrir fjóra
'Cod with saffron for four'

Þorskur með **basil/*basili**<dat> og sítrónu<dat>
'Cod with basil and lemon’

The datives of n.neut. names of countries and places are undisputed, the **-i** is unacceptable: *Austur-Tímori, *Bangladessi, *Bareini, *Belísi, *Brúneii, *Búrúndíi, *Chilei, *Djíbútíi, *Gíbraltari, *Japani, *Jemeni, *Kongói, *Mónakó, *Japani, *Afganistani, *Pakistani . . .

These are never mentioned in grammars.

The Database of Modern Icelandic Inflection/ Beygingarlýsing íslensks nútímamáls (BÍN)

271 thousand paradigms
5.8 million inflectional forms

Strong neuter nouns (-s genitive; other classifications exist)

- 36 inflectional classes
- 74,409 paradigms (including compounds)
- 4,216 base words (non-compounds)
- **3,165** relevant base words
- **383** words with the possibility of a dative in **-0**, 12.1%

12.1% of the strong neuter base words can have the dative ending **-0**. Enough to warrant a second look?

The inflectional classes in BÍN take all grammatical categories into account and all variants.

The Corpora: 554.5 million tokens

MÍM: **Mörkuð íslensk málheild** ‘The Tagged Icelandic Corpus’
25 million tokens, 21st C texts
The Árni Magnússon Institute for Icelandic Studies:
mim.arnastofnun.is

Osj: **Íslenskur orðasjóður** (Icelandic web pages)
500 million tokens, 2005, 2010
Leipzig University:
http://wortschatz.uni-leipzig.de/ws_ice/

Mbl: **Morgunblaðið** (Newspaper text)
29.5 million tokens, 2000-2007

The texts are tagged for word class and inflectional categories:
CombiTagger: MÍM and Íslenskur orðasjóður
IceNLP: Morgunblaðið

cf. malfong.is

The method

- Search for "n.neut.dat.sg. ending in **-i** or **-0**" in the corpora:
  - Word forms and frequency
  - All examples of each word form

- Sift through the word forms to exclude:
  - Errors in the PoS tags (in word class, gender, case, number, etc.)
  - Incomprehensible strings (errors, extraneous material)

- Analyze the word forms:
  - Lemmas
  - Morphological heads
  - Syllabic structure

Figures

	Dative word forms:		Examples:
	Org.neut.	Net.neut.	
	-i / -0	-i / -0	-i / -0
MÍM	18,664 / 392	12,375 / 229	320,552 / 4,432
Mbl	17,012 / 366	10,594 / 161	331,334 / 2,570
Osj	135,411 / 3,620	* / 1,641	6,806,074 / 53,325

Notes:
Org.neut.: Results from the search
Net.neut.: Actual dative forms, after removal of errors
*: The dative forms in -i in Osj. (Íslenskur orðasjóður) were too many to correct by hand.

Nb: The figures overlap as the same dative forms can occur in all sources. The total number of dative word forms in all the sources is not available.

An example word form

TIDNI:1|	Frequency of word form in source
OMYND:**bremsugúmmí**	Dative form ‘brake rubber’
|SKIL:bremsu_gúmmí	Compound boundary
|HAUS:gúmmí	Morphological head
|OGR:s	Morph. analysis (s = compound)
|ÞGF:0	Dative ending: (-i/-0)
|STOFNG-1:CúCCí	Structure of stem 1
|STOFNG-2:CVCCV	Structure of stem 2
|ATKV:2	Number of syllables
|HEIM:mbl	Source
|ATH:	Note

Morpholgical head **gúmmí** -i/-0: Mbl: 22/9 MÍM: 21/10

Number of syllables in morphological heads

No. of syllables	MÍM & MBL -0 / -i	=	%	Osj -0
1	227 / 10,765		2.1%	816
2	99 / 776		11.3%	528
3	47 / 319		12.8%	235
4	15 / 70		17.6%	69
5*	2 / 12		14.3%	27
6*	0 / 2		0%	0

* The analysis of 5 and 6 syllable stems is inconsistant and not statistically significant.

Examples of variants in non-compounds

	Mbl	MÍM	
	-i / -0	-i / -0	
badminton	4 / 34	3 / 15	‘badminton’
bíó	79 / 149	29 / 81	‘cinema’
granít	17 / 8	9 / 18	‘granite’
gúmmí	22 / 9	21 / 10	‘rubber’
lottó	10 / 2	4 / 16	‘lotto’
banjó	- / -	6 / 3	‘banjo’
beikon	6 / 2	0 / 0	‘bacon’
bingó	- / -	1 / 9	‘bingo’
glimmer	- / -	5 / 2	‘glitter’
íshokkí	- / -	3 / 8	‘ice hockey’
mangó	- / -	1 / 13	‘mango’
pestó	2 / 2	0 / 4*	‘pesto’
fé	**-**	**-**	**2 / 564** (+12 **-féi**, all sources)

*Pestó: 16 examples in MÍM; 3 (not dat.) are correctly tagged.

Problems

The number of errors in the tagging

> The automatic morphosyntactic tagging accuracy has been estimated as 88.1-95.1%, depending on text type. (Loftsson et al., 2010)

5-10% errors in our corpus: 27.7 – 55.4 million errors?

The dative in MÍM/Mbl:

40,064 word forms returned
24,900 word forms useable
62.2% net return

Data scarcity

klórít ‘chlorite’: 5 examples -i/-0: 8/2

Results

The dative -0 in strong neuter nouns appears in

- the classic exceptions: fé, hné, tré (universal, cf. the grammars)
- multisyllabic words of foreign origin: badminton, bíó, granít, gúmmí ... (variable, depending on stem structure*)
- the greater the number of syllables, the greater possibility of -0 (single syllable: 2.1%, 4 syllables: 17.6%)
- the names of countries and place names: Íran, Mexíkó (universal; not discussed here)

* Previous research indicates that phonotactics influence the choice between -0/-i.

Implications for standardization

- Is it justifiable to abstract and prescribe on partial data?
- Should standardization not be based on research?

Implications for grammatical theory?

- The traditional Icelandic grammars are partial. The creation of the MIM Corpus and other Icelandic text collections and corpora makes it possible to validate the descriptions of Icelandic inflection.

- "Inflectional classes are of particular interest, because they constitute a kind of autonomous morphological complexity which has no direct relationship to other levels of linguistic description, and hence there is no other objective way of assessing a theoretical characterisation of them."

  (Brown & Evans 2012)

- It would be of interest to us to use machine learning methods to validate models of the Icelandic inflectional system. The dative singular discussed here is very simple, compared to some other parts of the system. This is only a pilot project.

Thank you for your time!

Kristín B: kristinb@hi.is
Jón Friðrik: jfd1@hi.is

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

References

Brown, Dunstan, & Roger Evans. 2012. Morphological complexity and unsupervised learning: Validating Russian inflectional classes using high frequency data. Springer.

Guðrún Kvaran. 2005. *Orð*. Handbók um beygingar- og orðmyndunarfræði. [Words. Handbook of inflection and word formation.] Reykjavík, Almenna bókafélagið.)

Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir and Eiríkur Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. In *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the 7th International Conference on Language Resources and Evaluation (LREC 2010)*. Valetta, Malta.

Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. *LREC 2012 Proceedings:* Proceedings of "Language Technology for Normalization of Less-Resourced Languages", SaLTMiL 8-AfLaT 2012.[http://www.lrec-conf.org/proceedings/lrec2012/index.html: Workshops, SaLTMiL-AfLaT]

Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir & Hrafn Loftsson. 2012. The Tagged Icelandic Corpus (MÍM). *LREC 2012 Proceedings*: Proceedings of "Language Technology for Normalization of Less-Resourced Languages", SaLTMiL 8-AfLaT 2012. [http://www.lrec-conf.org/proceedings/lrec2012/index.html: Workshops, SaLTMiL-AfLaT]

Valtýr Guðmundsson. 1922. *Islandsk grammatik*. H. Hagerups forlag, København.

Quasthoff, Uwe, Sabine Fiedler, Erla Hallsteinsdóttir. 2012. Frequency Dictionary, Icelandic. Leipziger Universitätsverlag.

Links

Beygingarlýsing íslensks nútímamáls (BÍN) [The Database of Modern Icelandic Inflection] bin.arnastofnun.is

Mörkuð íslensk málheild (MÍM) [The Tagged Icelandic Corpus] mim.arnastofnun.is

Íslenskur orðasjóður. http://wortschatz.uni-leipzig.de/ws_ice/

CombiTagger: malfong.is

IceNLP: malfong.is

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM