já - er svarið

Beygingarlýsing íslensks nútímamáls

Stofnun Árna Magnússonar í íslenskum fræðum

The Database of Modern Icelandic Inflection

Kristín Bjarnadóttir

The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland

SaLTMIL-AfLat
Istanbul May 22th 2012

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

The Database of Modern Icelandic Inflection (DMII)

A database storing full inflectional forms

Topic:

- The aim in creating the DMII
- A short description of the DMII
- Why is the DMII not a productive system of rules?
  - The necessary information for a rule system was not available
  - A rule system would overgenerate wildly
- The problem of a large tag set: Data scarcity
- The two aspects of the DMII: LT and online

‘Modern’ in the title is used in the sense ‘current’.

The Database of Modern Icelandic Inflection

A database storing full inflectional forms

Used in Language Technology.
Accessible on-line for the general public.

Headwords/paradigms:	271,400
Inflectional forms (with tag):	5,800,000
Unique word forms:	2,800,000
PoS tag set:	700 +
Inflectional classes:	630 +

25% of the inflectional classes for the major word classes describe the idiosyncratic inflection of single words.

The aim in creating the DMII

Showing Icelandic inflectional ‘as is’

- All inflectional forms, including variants
- No overgeneration
- No overlap between inflectional classes, i.e., a word can only belong to 1 inflectional class (including variants)

- An inflectional class is a bundle of inflectional rules for a set of words.

- Production: A simple matrix of inflectional rules, with flags for values for inflectional categories: +/-singular, +/-plural, etc.

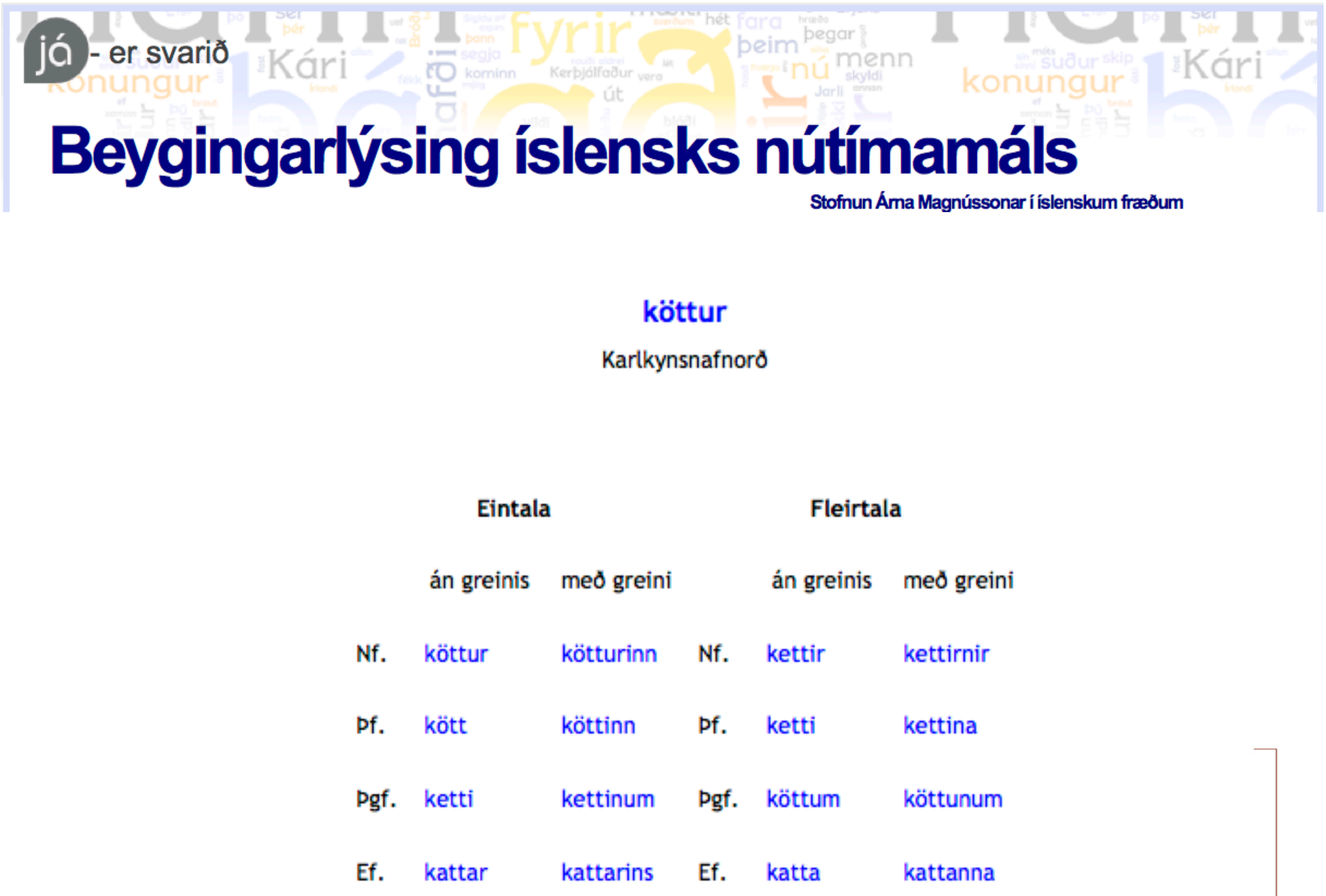
The Database of Modern Icelandic Inflection

Icelandic inflection is rich.

Inflectional categories for the major word classes and number of inflectional forms:

Nouns:	case (4), number (2), definiteness (2)	= **16**
Adjectives:	gender (3), case (4), number (2), definiteness (2), degree (3)	= **120**
Verbs:		= **106**
Finite:	voice (2), mood (2), tense (2), person (3), number (2)	= 48
Nonfinite:		= 58
	Imperative: voice (2), number (2) (+ 1 truncated form)	= 5
	Infinitive: voice (2)	= 2
	Past participle: gender (3), case (4), number (2), definiteness (2)	= 48
	Present participle:	= 1
	Supine: voice (2)	= 2

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM



Beygingarlýsing íslensks nútímamáls

Stofnun Árna Magnússonar í íslenskum fræðum

köttur

Karlkynsnafnorð

	Eintala			Fleirtala	
	án greinis	með greini		án greinis	með greini
Nf.	köttur	kötturinn	Nf.	kettir	kettirnir
Þf.	kött	köttinn	Þf.	ketti	kettina
Þgf.	ketti	kettinum	Þgf.	köttum	köttunum
Ef.	kattar	kattarins	Ef.	katta	kattanna

DMII: Output for LT projects

16 inflectional forms of the noun köttur ‘cat’

lemma id infl.form tag	English tag
köttur;416784;kk;alm;**köttur**;NFET	nom.sg.indef.
köttur;416784;kk;alm;**kötturinn**;NFETgr	nom.sg.def.
köttur;416784;kk;alm;**kött**;ÞFET	acc.sg.indef.
köttur;416784;kk;alm;**köttinn**;ÞFETgr	acc.sg.def.
köttur;416784;kk;alm;**ketti**;ÞGFET	dat.sg.indef.
köttur;416784;kk;alm;**kettinum**;ÞGFETgr	dat.sg.def.
köttur;416784;kk;alm;**kattar**;EFET	gen.sg.indef.
köttur;416784;kk;alm;**kattarins**;EFETgr	gen.sg.def.
köttur;416784;kk;alm;**kettir**;NFFT	nom.pl.indef.
köttur;416784;kk;alm;**kettirnir**;NFFTgr	nom.pl.def.
köttur;416784;kk;alm;**ketti**;ÞFFT	acc.pl.indef.
köttur;416784;kk;alm;**kettina**;ÞFFTgr	acc.pl.def.
köttur;416784;kk;alm;**köttum**;ÞGFFT	dat.pl.indef.
köttur;416784;kk;alm;**köttunum**;ÞGFFTgr	dat.pl.def.
köttur;416784;kk;alm;**katta**;EFFT	gen.pl.indef.
köttur;416784;kk;alm;**kattann**a;EFFTgr	gen.pl.def.

Creating the DMII

Inflectional class for the noun **köttur** ‘cat’
kk_sb_X.ur.-ar-ir.-i-inum
Set of stems: X1=**kött**, X2=**kött**, X3=**kett**, X4=**katt**

Matrix:

Nom.sg.	X1 +ur	Nom.sg.def.	X1 +urinn
Acc.sg.	X1 +	Acc.sg.def.	X1 +inn
Dat.sg.	X3 +i	Dat.sg.def.	X3 +inum
Gen.sg.	X4 +ar	Gen.sg.def.	X4 +arins
Nom.pl.	X3 +ir	Nom.pl.def.	X3 +irnir
Acc.pl.	X3 +i	Acc.pl.def.	X3 +ina
Dat.pl.	X2 +um	Dat.pl.def.	X2 +unum
Gen.pl.	X4 +a	Gen.pl.def.	X4 +anna

Creating the DMII

Inflectional class for the noun köttur ‘cat’

kk_sb_X.ur.-ar-ir.-i-inum

Set of stems: X1=**kött**, X2=**kött**, X3=**kett**, X4=**katt**

Paradigm:

Nom.sg.	kött+ur	Nom.sg.def.	kött+urinn
Acc.sg.	kött+	Acc.sg.def.	kött+inn
Dat.sg.	kett+i	Dat.sg.def.	kett+inum
Gen.sg.	katt+ar	Gen.sg.def.	katt+arins
Nom.pl.	kett+ir	Nom.pl.def.	kett+irnir
Acc.pl.	kett+i	Acc.pl.def.	kett+ina
Dat.pl.	kött+um	Dat.pl.def.	kött+unum
Gen.pl.	katt+a	Gen.pl.def.	katt+anna

The DMII: A multitude of variants

Inflectional classes: Over 630, due to inflectional variants

Examples of variants:
The genitive singular ending **–ar/–s** in masculine nouns:
þröskuldur ‘threshold’: gen. þröskuld**ar**/þröskuld**s**

The dative singular ending **–i/–0** in masculine and neuter nouns:

bátur ‘boat’ (masc.):	báti/bát
fennel ‘fennel’ (neut.):	fennel/fenneli
arsenik ‘arsenic’ (neut.):	arseniki/arsenik
ópíum ‘opium’ (neut.):	ópíum/*ópíumi

Name (3 variants pr. case):
Berglind (fem.): acc.+dat.: Berglind/Berglindi/Berglindu

Why not a productive system of rules?

Insufficient data

Lexicographic sources:

- Extensive vocabulary.
- Partial information on inflection, i.e., principle parts only:

  köttur (nom.sg.indef.), **kattar** (gen.sg.indef.), **kettir** (nom.pl.indef.)
- The rest of the paradigm is not (necessarily) predictable.

Grammatical surveys:

- Full paradigms of selected examples.
- A set of generalized inflectional classes, showing exceptions at times (often in notes).
- The emphasis is on the core Icelandic vocabulary, to the exclusion of loanwords, slang and informal language.
- Long history of research (1st survey 1651)
- Strong literary tradition; strong public interest, strong tendency to purism.

Generalization, as in the claim: “The dative ending –*i* is universal in neuter nouns”. →

A gap in the inflectional descriptions

Dative –i/-0 in multisyllabic neuter nouns

Hvönn er svolítið lík ***fennel****.*
‘Angelica is a little bit like fennel’
. . . *ásamt brytjuðu* ***fenneli****.*
‘. . . with diced fennel’
. . . *byrlað eitur, drepinn með* ***arsenik****.*
‘. . . poisoned with arsenic’
Hann var myrtur með ***arseniki****.*
‘He was murdered with arsenic’

cf.
Hann var myrtur með ***ópíum/*ópíumi****.*
‘He was murdered with opium’

Why not a productive system of rules: (2)

It would **overgenerate**. The data has to be collected first.

Research for the DMII

The search for possible variant inflectional forms

- All available sources are used:
- Archives, digitized text collections, corpora, Google (at a pinch)
- Native speaker intuition (linguists and others, including users on-line)
- The search is sometimes inconclusive:

Yggdrasill (masc.) ‘the great tree’ from Norse mythology:

- o Dative **Yggdrasil** or **Yggdrasli**?
- o The regular dative would be **Yggdrasli**.
- o No examples of the dative can be found in the Old Icelandic sources; modern examples are dubious.
- o The dative of the neuter noun **drasl** ‘rubbish’ is **drasli**.
- o When referring to a shop in today’s Reykjavík named **Yggdrasill** speakers seem to avoid the dative. (It makes people laugh.)
- o Examples of **Yggdrasli** appear on the web, mostly in disputes on the dative ...

Research for the DMII

Ambiguous inflectional forms make searching time-consuming

Inflectional forms in DMII	5,881,374	
Unambiguous	1,850,090	31.5%
Ambiguous within 1 lemma	3,619,482	61.5%
Ambiguous between lemmas	63,641	1.1%
Ambiguous within and between lemmas	348,161	5.9%

Inflectional form: Word form with inflectional tag.

Word forms (unique charater strings):	2.8 million
Unambiguous (unique inflectional forms):	1.8
Ambiguous:	1.0

Searching in unannotated text can be extremely slow.

Using the MÍM Corpus (The Tagged Icelandic Corpus)

cf. Sigrún Helgadóttir et. al., Poster session here at SaLTMIL

- 25 million running words from 21st Century texts.
- Compatible tags with the DMII.
- For the reason of ambiguity mentioned before, the first opportunity of systematic research on the inflectional system for the use in the DMII.

First stage of comparing the MÍM and the DMII:

Tokens in MÍM	16,245,429
Unique tagged forms in MÍM	737,856
In DMII	425,238
Not in DMII	312,618

Analysis of the word forms from the MÍM Corpus not found in the DMII

312,618 tagged word forms (or strings) (out of 737,856)

True Icelandic inflectional forms		60%
Miscellaneous strings		40%
Foreign words	25%	
Errors	6%	
Abbreviations & acronyms	1.7%	
Computer strings (Urls, etc.)	0.7%	

All word forms showing features of adjustment to the Icelandic inflectional system are counted as Icelandic inflectional forms.

Approx. additions to the DMII: 125,000 paradigms
+ uncounted “new” variants.

Data scarcity in a rich morphology

The MÍM Corpus:

** 623,000 inflectional forms.

The present DMII:

270,000 paradigms 5,800,000 inflectional forms

DMII with addions form MÍM:

395,000* paradigms 8,300,000* inflectional forms

****Figure for lemmas in MÍM not available yet.*

** Estimated figures*

- This comes as no surprise to Icelandic lexicographers; we have always had to cope with the problem of scarcity.
- A description of Icelandic inflection based solely on the MÍM corpus would be very meager.

Conclusion

- The DMII was initially made for two purposes:

  As an LT resource (including lexicography)

  For reference for the general public

- In spite of a long history of research, the available data was insufficient. The DMII has therefore become increasingly important in research on Icelandic morphology.

- Both corpora and lexicographic data is needed for the DMII.

- As the scope of the DMII expands, the production of a rule system becompes more feasible, if one is needed.

With a PS on availability … →

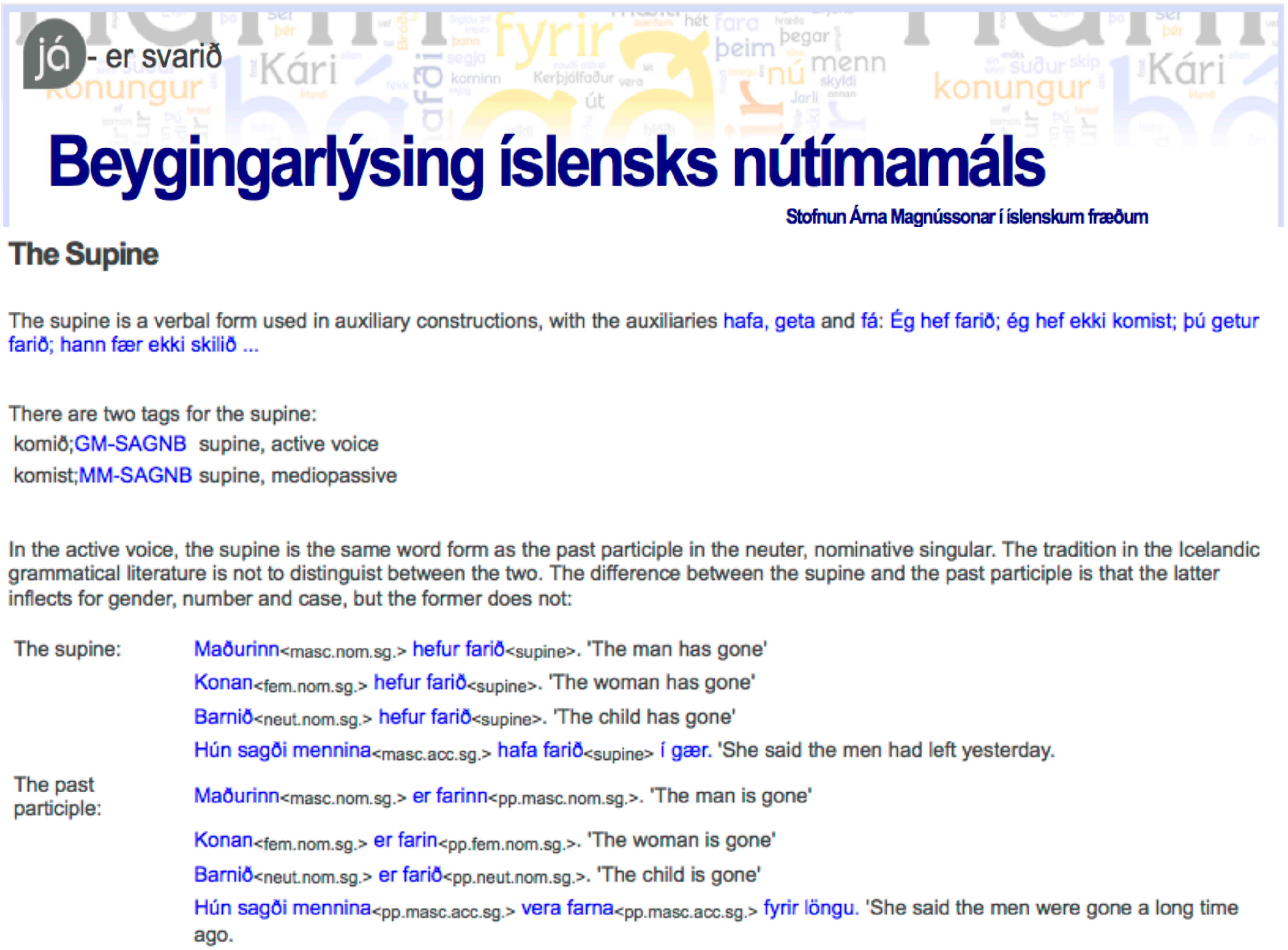
Availability

The DMII is available for LT projects, free of charge.
Download: http://bin.arnastofnun.is/gogn/
Online version: http://bin.arnastofnun.is

The website is as yet only in Icelandic. Send me an email for an English version of the conditions on the use of data from the DMII or any questions: **kristinb@hi.is**

The website is being restructured. The new website will contain extensive comments, explanations of grammatical features, etc. Work on an English version of this part of the site is in progress.

The metalanguage of the paradigms themselves will be Icelandic.
Be in touch if you need information.



já - er svarið

Beygingarlýsing íslensks nútímamáls

Stofnun Árna Magnússonar í íslenskum fræðum

The Supine

The supine is a verbal form used in auxiliary constructions, with the auxiliaries hafa, geta and fá: Ég hef farið; ég hef ekki komist; þú getur farið; hann fær ekki skilið ...

There are two tags for the supine:

komið;GM-SAGNB	supine, active voice
komist;MM-SAGNB	supine, mediopassive

In the active voice, the supine is the same word form as the past participle in the neuter, nominative singular. The tradition in the Icelandic grammatical literature is not to distinguist between the two. The difference between the supine and the past participle is that the latter inflects for gender, number and case, but the former does not:

The supine:	Maðurinn<masc.nom.sg.> hefur farið<supine>. 'The man has gone'
	Konan<fem.nom.sg.> hefur farið<supine>. 'The woman has gone'
	Barnið<neut.nom.sg.> hefur farið<supine>. 'The child has gone'
	Hún sagði mennina<masc.acc.sg.> hafa farið<supine> í gær. 'She said the men had left yesterday.
The past participle:	Maðurinn<masc.nom.sg.> er farinn<pp.masc.nom.sg.>. 'The man is gone'
	Konan<fem.nom.sg.> er farin<pp.fem.nom.sg.>. 'The woman is gone'
	Barnið<neut.nom.sg.> er farið<pp.neut.nom.sg.>. 'The child is gone'
	Hún sagði mennina<pp.masc.acc.sg.> vera farna<pp.masc.acc.sg.> fyrir löngu. 'She said the men were gone a long time ago.

Thank you for your attention

Kristín B
kristinb@hi.is

SaLTMIL-AfLat
Istanbul May 22th 2012