Online Data on Icelandic Inflection. Descriptive or Prescriptive: "Why, for whom, by whom" and how?

Kristín Bjarnadóttir & Kristín Ingibjörg Hlynsdóttir

The topic of this paper is the production of a prescriptive subset of the Database of Modern Icelandic Inflection, which was originally designed for use in language technology and is descriptive in nature. The data has been available to the general public online from 2004, and users have expressed the need for prescriptive data. Such data is also needed in some language technology projects, such as spell checking and language production of any kind. The method of classification used to produce the prescriptive subset, the DMII Core, is described.

1. Introduction

The Database of Modern Icelandic Inflection (DMII) contains paradigms of Icelandic words. The DMII was originally produced for use in language technology (LT) and lexicography, but the data was also to be made available to the general public online. Work on the DMII started in 2002 and the first version of the website was opened in 2004. The original DMII was descriptive in nature (cf. Bjarnadóttir 2012, Bjarnadóttir et al. 2019).[1]

In 2017, a grant from the Icelandic Language Technology Fund (Máltæknisjóður) made restructuring of the original DMII database possible. A part of the restructuring is classifying the data according to acceptability, with reference to Icelandic language standards. The result of this work is an enhancement of the downloadable LT data, and the DMII Core, which is a prescriptive subset of the DMII intended for learners of Icelandic. The topic of this paper is the creation of the DMII Core, specifically the classification system used to mark the prescriptive data.

The paper is structured as follows: Section 2 contains a description of the DMII and the need for prescriptive data. Section 3 contains a description of variants in Icelandic morphology. Section 4 briefly describes Icelandic standards of language use. Section 5 contains a description of the DMII Core. Section 6 describes the classification system. The conclusion is in Section 7.

[1] The work described here was done at The Árni Magnússon Institute of Icelandic Studies by the authors of this paper.

2. DMII and the Need for Prescriptive Data

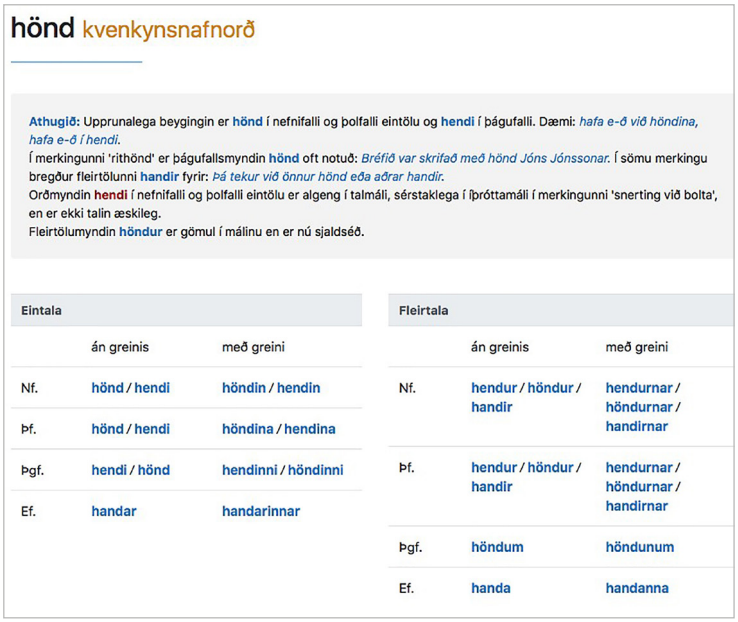
The DMII data is accessible online at the website of The Árni Magnússon Institute for Icelandic studies[2], and it is also downloadable for use in LT. At the time of writing, DMII contains over 292,000 headwords and over 6.2 million inflectional forms. The online DMII is extensively used by the Icelandic public as a reference on inflection, and it has been popular from the start in 2004. In the year starting on June 1 2018, there were 201,207 users and 1,714,726 pageviews, and 9% growth from the previous year.

From the start, the DMII was to be a description of Icelandic inflection "as is", i.e., without regard to correct spelling, grammar, vocabulary, etc. For LT analysis, the key factor is inclusive data, showing all varieties, ranging from context-bound obsolete word forms to non-standard inflectional variants and non-standard compound formations and words. As clearly stated on the website, the DMII was from the beginning meant to be descriptive, and it was neither to be an exhaustive source of Icelandic vocabulary nor an authority on the acceptability of vocabulary, inflectional variants or spelling. In spite of rather forcefully phrased instructions to the contrary, online users still expect online paradigms published by an accredited authority on the Icelandic language, such as the Árni Magnússon Institute for Icelandic Studies, to be "correct", i.e., prescriptive. The expectation is that all word forms appearing on the website are "best usage", all spelling variants are "good", and all words are acceptable. Some users even expect the DMII to be exhaustive, assuming that the absence of a word indicates that it is not an acceptable Icelandic word.

To assist the users of the original website, the paradigms include usage notes. These explain the choice between inflectional variants, such as constrictions of the use of individual inflectional forms. The notes are in Icelandic only, and they make the website unsuited for any but native or near-native speakers of Icelandic, as the users have to make the final choice between variants themselves. Fig. 1 shows a paradigm with notes.

The problems learners of the language face when using the online DMII are not trivial. The data is both too complex, as when addressing instances of obscure usage, and not instructive enough, in not presenting simple information on acceptability. The paradigm for *hönd* is a case in point, as the inflectional variants are acceptable to different degrees.

[2] www.arnastofnun.is/en



hönd kvenkynsnafnorð

Athugið: Upprunalega beygingin er **hönd** í nefnifalli og þolfalli eintölu og **hendi** í þágufalli. Dæmi: *hafa e-ð við höndina, hafa e-ð í hendi.*
Í merkingunni 'rithönd' er þágufallsmyndin **hönd** oft notuð: *Bréfið var skrifað með hönd Jóns Jónssonar.* Í sömu merkingu bregður fleirtölunni **handir** fyrir: *Þá tekur við önnur hönd eða aðrar handir.*
Orðmyndin **hendi** í nefnifalli og þolfalli eintölu er algeng í talmáli, sérstaklega í íþróttamáli í merkingunni 'snerting við bolta', en er ekki talin æskileg.
Fleirtölumyndin **höndur** er gömul í málinu en er nú sjaldséð.

Eintala	án greinis	með greini
Nf.	hönd / hendi	höndin / hendin
Þf.	hönd / hendi	höndina / hendina
Þgf.	hendi / hönd	hendinni / höndinni
Ef.	handar	handarinnar

Fleirtala	án greinis	með greini
Nf.	hendur / höndur / handir	hendurnar / höndurnar / handirnar
Þf.	hendur / höndur / handir	hendurnar / höndurnar / handirnar
Þgf.	höndum	höndunum
Ef.	handa	handanna

Figure 1: The paradigm for *hönd* 'hand' on the DMII Website.[3]

3. The DMII and Icelandic Morphology

The morphology of Icelandic is rich, both in inflection and word formation, and variant inflectional forms are possible for very many inflectional categories. The number of inflectional variants in each pattern varies, from one set as in the noun *bátur* 'boat' where the dative singular is *báti/bát*, to almost a full set of duplicate inflectional forms in the adjective *þögull* 'silent', e.g., *þögulan/þöglan* (masc.acc.sg.indef.), *þögulum/þöglum* (dat.sg.indef.), etc. There are 56,143 sets of inflectional variants in the database, most of which contain two variants, as in *báti/bát*, with three variants found in only 129 sets, as is the case for *hönd*,

[3] Translation of the note ('Athugið'): The original inflection is nom. & acc.sg. **hönd**, dat. sg. **hendi**, e.g. *hafa eitthvað við höndina* 'have sth. at hand', *hafa e-ð í hendi* 'have sth. in hand' (i.e., 'be sure of sth.'). In the the sense 'handwriting' the dat.sg. **hönd** is common: *Bréfið var skrifað með hönd Jóns Jónssonar* 'The letter was written in JJ's handwriting. In the same sense, the plural **handir** is also found, e.g. *Þá tekur við önnur hönd eða aðrar handir* 'Then we have another scribe or scribes.' The variant **hendi** in nom. & acc. sg. is common in colloquial language, especially in sports in the sense 'a touch of a hand with a ball'. This is not considered good usage. The plural **höndur** is old-fashioned and now very rare.

e.g., *hendur/höndur/handir*. Two variants of the same inflectional category can be equally valid in (almost) any context, as in *báti/bát*, but the choice between variants can also be context-sensitive or semantically restricted, as in the dative singular *hönd* in the meaning 'handwriting', where the unmarked dative is *hendi*, as shown in Fig. 1.

4. Icelandic Standards for Language

Icelandic language standards appear in the "Rules of spelling and punctuation" (*Ritreglur* (2016) & *Reglur um greinamerkjasetningu* (2018)), and in "The Dictionary of Spelling" (*Stafsetningarorðabókin*) (Sigtryggsson 2016). The work on the prescriptive version of the DMII starts with these standard references.

There is no comprehensive standard for Icelandic inflection. Icelandic dictionaries show only a small subset of inflectional forms, i.e., nom.sg. & pl., and gen.sg. (3 forms out of 16), and for verbs the principal parts are shown (4-5 out of over a hundred). The remainder of the paradigms are not always deductible from the available forms. Grammar books also contain fragmentary descriptions, as the inflectional system is described "top-down", with few examples. The result is that producing the paradigms for the DMII involved considerable research where the standards were of no help.

5. The DMII Core

The DMII Core is a subset from the DMII, designed to meet users' demand for data from the DMII in a prescriptive context, especially for use in schools and for learners of the language. It is created for third party publication, and it is accessible through a RESTful API,[4] open for everyone. The API allows users to send simple queries and receive full paradigms in JSON-format as a response. In order to make the transition from descriptive to prescriptive, the core vocabulary of the DMII has been checked against the standards of good usage mentioned above.

The DMII Core contains a subset of the vocabulary of current Icelandic, i.e., common non-domain specific words, and a selection of named Icelandic entities, i.e., personal names, common place names, etc. The spelling is standardized. The sources of the vocabulary are *Íslensk nútímamálsorðabók* [*The Modern Icelandic Dictionary*] (Jónsdóttir & Úlfarsdóttir 2018), containing approximately 50,000 headwords, with additions from the 50,000 most frequent lemmas in The Icelandic

[4] API: application program interface. REST: Representational State Transfer, an architectural pattern for creating web services. A RESTful service implements that pattern.

Gigaword Corpus (Steingrímsson et al., 2018). The overlap between these two sources is great, and the total number of paradigms in the DMII Core now stands at 58,086, out of 293,783 in the DMII.

The paradigms in the DMII Core have been simplified as much as possible, without omitting equally valid variants, showing only the best forms, or the variants not limited by specific usage restrictions. The classification system created for this use is described in Section 6.

6. The Classification System

The classification system used to create the DMII Core extends to both the vocabulary and individual inflectional forms. It was created by analysing all the handcrafted notes on individual paradigms in the DMII as they stood in March 2018 when work on a new version of the DMII was started. The results were then formalized, both with regard to the Icelandic standard (as for spelling, word formation variants, and inflectional variants), usage considerations, the grammatical features involved, and considerations of style and/or register.

The classification system is in three parts, two of which apply to both words and variants, i.e., genre/register (Section 6.1) and correctness grade (Section 6.3). Variants are also graded according to their relative value (Section 6.2). The parts of the classification system are interdependent, especially the correctness grade which is largely estimated on the basis of the other three.

6.1. Genre or Register

Genre or register is used to sort words and inflectional forms according to style, age and other comparable features. The classification functions as a tool for finding the relevant vocabulary for inclusion in the Core, omitting marginal data. The division in genres/registers is as follows:

- Included in the DMII Core: derogatory, obscene, informal (and unmarked words and inflectional variants).
- Not included in the DMII Core: rare, old-fashioned, obsolete; poetic language, regional; unadapted loan words.

6.2. Value of Inflectional Variants

As stated above, there are 56,143 sets of inflectional variants in the DMII. Most of the sets contain two variants each, but 129 sets contain three variants. The variants have been evaluated and graded, to indicate their status with respect to

the other variant in each set. The classification also includes the marking of inflectional forms only found in fixed expressions or idioms. The value determines inclusion or exclusion in the Core, and they are as follows:

- Equal: There are no restrictions on the use of either variant in a set. Example: The two genitive singular forms of the masculine noun *bekkur* 'bench', *bekkjar/bekks*, are equally valid, according to *The Dictionary of Spelling*. Both are included in the DMII Core.
- Dominant/yielding: One variant in a set is acceptable without restriction, the other one is not. Example: The masculine noun *refur* 'fox' has sets of two variants in the nominative and accusative plural: *refir/refar* (nom. indef.), *refirnir/refarnir* (nom.def.), and *refi/refa* (acc.indef.), *refina/refana* (acc. def.). The first variant in each set is dominant and thus included in The DMII Core. The other variants are yielding, as they are obsolete in Modern Icelandic and only used in idioms such as *til þess eru refarnir skornir* 'that's why the fox are cut' (i.e., 'that is why a certain action was taken'). These are omitted from the DMII Core.
- Obsolete inflectional forms, mostly found in idioms or fixed expressions. Example: *skjöldu* (acc.pl.) of the masculine noun *skjöldur* 'shield', used in a number of idioms. The corresponding modern inflectional variant is *skildi*. These are omitted from the DMII Core.

Frequency is used to determine the value of variants, when the standards are of no help and the distribution seems to be independent of context.

6.3. Correctness Grading of Words and Inflectional Variants

The Correctness grade is used in the work on the Core to mark a word's or variant's correctness according to prescriptive grammar rules and standardized spelling. Grades range from 0 to 4. The only grade included in the DMII Core is "1". It should be noted that the equal value of variants is quite common and in that case both forms are assigned the value "1" and included in the DMII Core. The grading is as follows:

1) Not applicable, dependent on genre or register, cf. Section 5.1.
2) Correct, default value, most words and inflection forms.
3) Used, not as good. Mostly used of common words and inflectional forms not approved of in the standards.
4) Not good. Used of words and inflectional form directly in contradiction to the recommendations of the standards.
5) Very bad. Used of words and inflectional forms deemed errors in the standards.

In the production of the DMII Core, the Icelandic standards are used to determine the grade of a word or inflectional variant according to correctness, as far as possible. The goal is to find the items to mark with "1". One problem is that the data in the standards is simply too scarce to answer all questions when it comes to individual words or inflectional forms. The *Rules of spelling and punctuation* are generalizations, with relatively few examples, and they are not always easy to interpret or extrapolate. The problem of scarcity also applies to T*he Dictionary of Spelling*, as the vocabulary is small.

6.4. Use of the Classification System

The aim in the creation of the DMII Core is to produce a prescriptive, simplified version of the inflectional paradigms, containing only the most common (and correct) variants, as in the 2nd column in Table 1.

	Core	A	DMII	A	B	C
Nom.sg.	*hönd*	1	*hendi*	3	Informal	
Nom.sg.def.	*höndin*	1	*hendin*	3	Informal	
Acc.sg.	*hönd*	1	*hendi*	3	Informal	
Acc.sg.def.	*höndina*	1	*hendina*	3	Informal	
Dat.sg.	*hendi*	1	*hönd*	0		Meaning
Dat.sg.def.	*hendinni*	1	*höndinni*	0		Meaning
Gen.sg.	*handar*	1				
Gen.sg.def.	*handarinnar*	1				
Nom.pl.	*hendur*	1	*höndur*	0	Old-fash.	Yielding
			handir		Old-fash.	Yielding
...						

Table 1: Grading: A: Correctness grade, B: Genre/register, C: Value.

The difference between the DMII and the Core (cf. the paradigm for *hönd* in Section 2) is demonstrated in Table 1. The column headed 'Core' only contains acceptable variants, i.e., Correctness grade 1 in the 3rd column (A), unmarked for genre, register or value. The DMII contains the complete set of data, including Correctness grade 0 or 3 (A), and variants marked for genre, register or value (B & C). The omission of a variant does not necessarily imply that it is wrong, and for a full exposition of variants the users are referred to the DMII Website.

7. Conclusion

The classification system described above has been used to mark all sets of inflectional variants currently in the DMII database, and all headwords in the DMII Core. The simplified DMII Core is made for language learners and for use in teaching material. This does not in any way detract from the importance of access to the complete set of data on the DMII Website. In turn, work on the DMII Core has led to better usage notes on the website, with more information on prescriptive data found in the classification system, with extensive cross-referencing of less than optimal instances of usage to standardized forms.

The DMII Core API is open for everyone to use. It is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License (CC BY-ND 4.0).

Acknowledgements

The authors wish to thank the Icelandic Language Technology Fund for the grant which made the restructuring of the DMII possible. Thanks are also due to our DMII colleagues Samúel Þórisson, Steinþór Steingrímsson and Trausti Dagsson, and to the editors of the D*ictionary of Spelling*, Jóhannes B. Sigtryggsson, and the *Dictionary of Modern Icelandic*, Halldóra Jónsdóttir and Þórdís Úlfarsdóttir, for being very generous with their time.

References

Bjarnadóttir, Kristín (2012): The Database of Modern Icelandic Inflection. In: *LREC 2012 Proceedings: Proceedings of "Language Technology for Normalization of Less-Resourced Languages", SaLTMiL 8 -- AfLaT 2012*, pp. 13-18.

Bjarnadóttir, Kristín, Kristín Ingibjörg Hlynsdóttir & Steinþór Steingrímsson (2019): DIM: The Database of Icelandic Morphology. In: P*roceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 146-154.

DMII = Kristín Bjarnadóttir (ed.): *Database of Modern Icelandic Inflection* [*Beygingarlýsing íslensks nútímamáls*]. Reykjavík: The Árni Magnússon Institute for Icelandic Studies. <www.bin.arnastofnun.is>

Jónsdóttir, Halldóra, & Þórdís Úlfarsdóttir (eds.) (2018): *Íslensk nútímamálsorðabók* Reykjavík: The Árni Magnússon Institute for Icelandic Studies. <islenskordabok.arnastofnun.is>

Ritreglur [Rules on Spelling] (2016). Published by The Ministry for Culture and Education, Reykjavík.

Reglur um greinarmerkjasetningu [Rules on Punctuation] (2018). Published by The Ministry for Culture and Education, Reykjavík.

Sigtryggsson, Jóhannes Bjarni (ed.) (2016): *Stafsetningarorðabókin* [The Dictionary of Spelling], 2nd ed. Reykjavík: The Árni Magnússon Institute for Icelandic Studies.

Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson & Jón Guðnason (2018): Risamálheild: A Very Large Icelandic Text Corpus. In: *Proceedings of LREC 2018*, pp. 4361-4366.

Kristín Bjarnadóttir
associate Research Professor
The Árni Magnússon Institute
for Icelandic Studies
Reykjavík
kristinb@hi.is

Kristín Ingibjörg Hlynsdóttir
researcher
The Árni Magnússon Institute
for Icelandic Studies
Reykjavík
kih4@hi.is