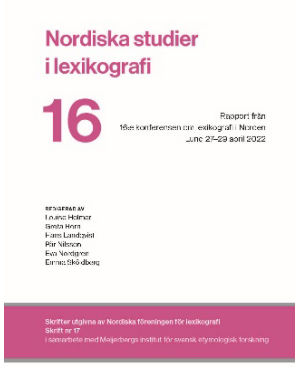
NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	When the users jump to conclusions. Presenting prescriptive information
Forfatter:	Kristín Ingibjörg Hlynsdóttir & Kristín Bjarnadóttir
Kilde:	Nordiska studier i lexikografi 16, 2023, s. 141–151
URL:	https://tidsskrift.dk/nsil/issue/archive



When the users jump to conclusions. Presenting prescriptive information

Kristín Ingibjörg Hlynsdóttir & Kristín Bjarnadóttir

The topic of this paper is a method of presenting acceptability to the users of the online version of the Database of Modern Icelandic Inflection (DMII), with a short description of the classification used and a reference to a survey of one week of online queries, a total of 117,685 searches. The DMII was originally descriptive, and the inclusion of non-standard inflectional and spelling variants is known to confuse users who expect prescriptive data. Prescriptive information is provided in usage notes presented with the paradigms, but the users are apt to stop at the search list and jump to conclusions on acceptability without reading the notes. Non-standard headwords therefore need to be marked in the search list itself, with cross references to the standard forms, as needed.

KEYWORDS: morphology, inflection, Icelandic, language standard, language technology resource

1. Introduction

The DMII is an online reference for the general public and a resource for language technology (LT). The project has been ongoing at the Árni Magnússon Institute for Icelandic Studies (AMI) since 2002. The website (bin.arnastofnun.is) shows full paradigms of over 333,000 headwords and the data is available as downloadable CSV files. A smaller prescriptive version is available through an application programming interface (API), The DMII Core (Bjarnadóttir & Hlynsdóttir 2020). The DMII is an important resource for Icelandic LT and the website is very popular among the general public, with over 7 million page views in 2021.

The DMII was originally descriptive and the purpose was to show language use "as is", i.e., both standard and non-standard usage, with LT use in mind. The inclusion of non-standard inflectional and spelling variants can, at times, confuse the users of the website, as they expect a source from AMI to show only what is correct, i.e. prescriptive data. This was coun-

tered by adding usage notes with paradigms, mostly to guide users when choosing between inflectional variants.

In 2019, a new version of the DMII was released with extended usage analysis. A new grading and classification system made it possible to add more non-standard forms and to grade and differentiate between standard and non-standard forms, to create the DMII Core and for other LT uses. As a result, many more non-standard forms are now displayed on the website, both inflectional variants and headwords. This has led to more usage notes being added to the paradigms, which partly solves the problem of guiding the users as to good usage, i.e., in the choice of variant inflectional forms within the individual paradigm. The choice between headwords needs to be addressed in a different way, as non-standard headwords need to be marked in the search list, with a cross reference to the standard form. This is doubly important, as the users tend to forget the descriptive nature of the DMII and regard all the data therein as correct or acceptable. Users are also apt to stop at the search list and when they do so they never discover the usage notes in the paradigms, i.e., they jump to conclusions about acceptability.

The topic of this paper is a method of presenting acceptability to the users as efficiently as possible, with a short description of the classification used and a reference to a survey of one week of online queries, a total of 117,685 searches.

2. Descriptive vs. prescriptive

Over 20 years ago, the original purpose of the DMII was use in LT, at a very early stage of that field in Iceland. As Icelandic is a heavily inflected language, the immediate need was for data for search engines, etc., containing as large a vocabulary as possible with all corresponding inflectional forms. The first version of the DMII, published in 2004, contained an average of 27 inflectional forms per paradigm. Inclusiveness was also an important feature, which is why the DMII had to be descriptive and not prescriptive. The DMII data includes the good, the 'not-so-good' and the downright erroneous, according to the Icelandic language standard.

The first version of the data contained no classification of acceptability, and broadly speaking, its main function was to link lemmas and inflectional forms for use in LT. The first online version was a side product to

the LT data, but it has gained in importance and it is now used extensively by the public. The online users' need for prescriptive data is unquestionable, and the same applies to today's LT uses, such as spell checkers, grammar checkers, and any kind of language production, such as translation services, query systems, etc. The gradual change of the DMII from purely descriptive data to prescription with a reference to the Icelandic language standards is described in NSL 15 (Bjarnadóttir & Hlynsdóttir 2020). The development of the DMII is described in detail on the DMII website.

3. The original search results

Headwords in the DMII are presented on the web with full paradigms and usage notes, based on the classification described in NSL 15 (Bjarnadóttir & Hlynsdóttir 2020). Headwords can be searched for using a search bar on the website and if the search string returns multiple headwords, the results are listed as shown in figure 1, as headword, word class and domain.

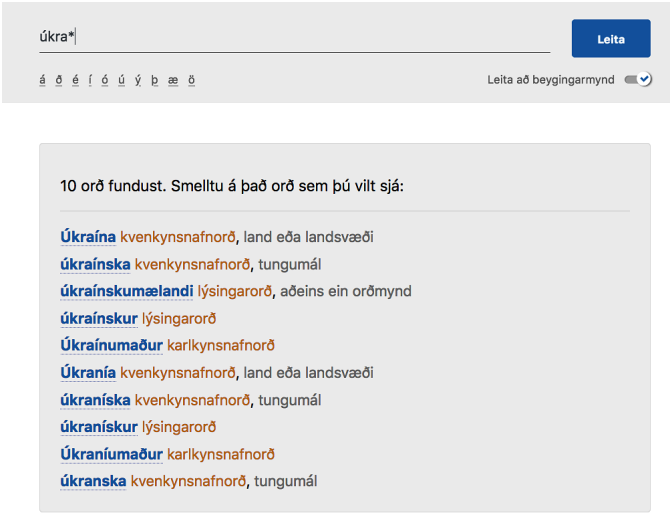
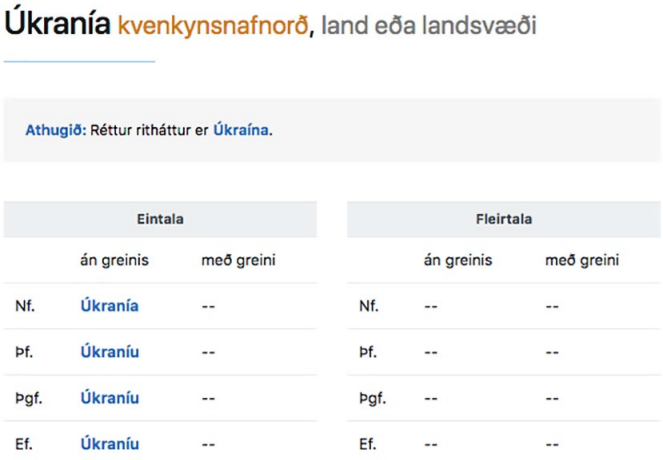


FIGURE 1. Searching for *úkra** in the online DMII.

Úkraína is the standard Icelandic form of the country name *Ukraine* but an alternative non-standard spelling variant is *Úkranía*. The headwords in figure 1 are all derived from these two variants, but the search result shows no indication of acceptability and the user might think that the variants *Úkraína* and *Úkranía* are equally acceptable, as they are both included in the

DMII. The user needs to click on a headword to see the paradigm in order to read the accompanying note to know whether the word form is correct or not. Selecting the word *Úkranía* from the list in figure 1 shows the paradigm, with a usage note saying that the correct spelling is *Úkraína*, cf. figure 2.



Úkranía kvenkynsnafnorð, land eða landsvæði

Athugið: Réttur ritháttur er **Úkraína**.

	Eintala		Fleirtala		
	án greinis	með greini		án greinis	með greini
Nf.	Úkranía	--	Nf.	--	--
Þf.	Úkraníu	--	Þf.	--	--
Þgf.	Úkraníu	--	Þgf.	--	--
Ef.	Úkraníu	--	Ef.	--	--

FIGURE 2. The paradigm of the word *Úkranía* with the usage note: "The correct spelling is ***Úkraína***."

Only selected parts of the classification for correctness produced for LT are displayed on the web as some of this data is specific to LT tasks and not suited for use by the general public. The classification is used as a tool to help create the usage notes which can be as long as needed as there is no need to save space. The notes can also contain references, i.e., links to further explanations. The aim is to make the notes as clear and readable as possible. The problem remains that many users only go as far as the search results, even though the explanations on the front page of the website clearly state the descriptive nature of the DMII.

The original source for headwords in the DMII was mostly lexicographical material, containing common non-standard word forms and spelling. With recent additions from the Gigaword Corpus (Steingrímsson et al. 2018) and work on error analysis, erroneous forms are deliberately added to the database with corrections to make sure to cover all common word forms for LT uses. This new type of data is of two types, i.e., errors in individual word forms (referred to as *non-standard word forms*)

and errors encompassing whole paradigms (referred to as *paradigms of errors*). These errors will be searchable with links to the correct forms online. It is important to make sure the users get a clear message that they are being referenced to a new word, so that they can see that the original form in their search string is not the standard form.

4. Learning from user search strings

Work on the DMII has been more LT focused in the last couple of years, with recent additions being produced as part of the Language Technology Programme for Icelandic (Almannarómur 2022). In order to make better use of the new data on the DMII website, the actual queries on the website were reviewed with the aim of finding out what the users were really searching for. All search queries from the week Jan. 24–30 2022 were collected, a total of 117,685 searches; 34,186 unique strings, with 12,021 strings not found in the DMII.

Many of the search strings not found in the DMII were ordinary, acceptable words missing from the DMII. As a result, approx. 2,500 new headwords were added to the database, using the Gigaword Corpus (Steingrímsson et al. 2018) for reference, and also adding some related compounds found in the Gigaword Corpus but not in the search list.

The remainder of the strings not found in the DMII contained various kinds of errors. A large portion of them were multiword search strings, such as:

- Noun phrases: *rauður hestur* 'red horse', *tveir ungir menn* 'two young men'
- Verbs with infinitive marker: *að gráta* 'to cry'
- Particle verbs: *ráðast á* 'attack'
- Prepositional phrases: *til upplýsingar* 'for information'
- Grammatical features included in the search string: *miðstig gamla* 'comparative old'
- Miscellaneous multiword strings: *ostur með sinnepi* 'cheese with mustard'

Other error strings contained wrong character sets, foreign queries, non-alphabetic characters and symbols, etc.

The remainder were real language errors, i.e., recognizable Icelandic word forms containing errors in spelling, word formation, typos, etc. These were analysed and classified according to the previously established system and then added to the DMII, as full-scale paradigms visible on the web or paradigms of errors and non-standard word forms. In the case of paradigms of errors and non-standard word forms, the corresponding correct headword and paradigm was sometimes missing from the DMII and had to be added.

The analysis of the search strings for words already in the DMII gave indications of the purpose of the search, which usually seemed to be for inflection, spelling or even word formation, although some of the strings are a bit harder to interpret. In the case of inflection, the users need to access the full paradigm, but in other cases the users may decide to make do with the search list, which means they will not see the needed notes on acceptability. Analysing the search strings gives limited scope for interpretation, and doing a thorough user survey would be very interesting. The simple analysis of the search strings described here does, however, confirm the need for a clearer presentation of the standard spelling in the DMII and the importance of including as many headwords as possible, including non-standard ones. The key issue is making it as easy as possible for the users to access the information.

5. Changes to the presentation of non-standard forms

As previously stated, many users seem to believe that everything found on the website is correct because they expect the data to be prescriptive. This is known from feedback given in e-mails, on social media, etc. The DMII describes the language "as is" and the scope of the DMII is much larger than any part of the Icelandic language standard. The DMII is, however, anything but exhaustive, either in vocabulary or inflectional forms. Some users have misconceptions about words not found in the DMII and they assume that words missing from the database have been deemed incorrect by the editors. The users have a tendency to regard the DMII as a language standard for Icelandic, which it is not. Some readers even expect the DMII to be exhaustive and assume missing words to be nonexistent in the language, which is certainly not the case.

The DMII editorial concept has been that the users must be able to find non-standard words and inflectional forms and the aim is also to explain

why they are not considered acceptable, as far as possible. The vocabulary of the DMII (or any other source) is not exhaustive and it never will be, but the goal is to include as much as possible.

5.1. Search heads

After making sure the users find what they are looking for, the next task is ensuring that they actually understand the given information, preventing them from jumping to conclusions. As shown in figure 1 in section 3, search lists were misleading for users that did not proceed to the usage notes. Adding data in a shorter form (search heads) to the search lists themselves solves this problem. The search heads are standardised based on style/register and grade but semi-manually added to each word. The headers also contain word class and domains.

Examples of search heads are shown in the following tables:

TABLE 1. Corrections, (usually) showing target word or word form.

Úkranía kvenkynsnafnorð. Réttur ritháttur er **Úkraína**.	[Correct form]
Egilssaga kvenkynsnafnorð. Hefðbundinn ritháttur er **Egils saga**.	[Traditional form]
Hercules karlkynsnafnorð. Íslenski ritátturinn er **Herkúles**.	[Icelandic form]
ábrestur kvenkynsnafnorð. Afbrigði af **ábrystir**.	[Variant form]
kólumbín hvorugkynsnafnorð. Eldra heiti á **níóbín**.	[Older term]
kjurr lýsingarorð. Framburðarmynd af **kyrr**.	[Pronunciation form]
hör hvorugkyn. Rétt er að hafa orðið í karlkyni.	[The correct gender is masculine]

TABLE 2. Words that are not fully acceptable, without direct reference to a standard form.

sinnhver óákveðið fornafn. Ekki viðurkennt mál.	[Unacceptable]
selebreita sagnorð. Sletta.	[Unacceptable loanword]
Discorites karlkynsnafnorð. Erlendur ritháttur.	[Foreign spelling]

The search heads are also used for references between equally correct forms or sets of easily confused words, such as homophones.

TABLE 3. Homophones.

skrýtinn lýsingarorð. Einnig ritað **skrítinn**.	[Also written]
híði hvorugkynsnafnorð. Orðið **hýði** hefur aðra merkingu.	[Confusion set]

Style or register is simplified for presentation in the search list, using "Gamalt" 'old' when the style is obsolete, Old Icelandic or old-fashioned, and "Sjaldséð" 'rare' for dialectal, poetic or rare words. The actual usage notes with the paradigms usually contain more specified data, and these are individually written for each paradigm.

TABLE 4. Age, style or register.

afbatan kvenkynsnafnorð. Gamalt	[Old (classified obsolete)]
mánadagur karlkynsnafnorð. Gamalt	[Old (classified Old Icelandic)]
bedraga bedró bedrógum bedregið sagnorð. Gamalt	[Old (classified old-fashioned)]
aflaslór hvorugkynsnafnorð. Sjaldséð	[Rare (classified dialectal]
aldurlok hvorugkynsnafnorð. Sjaldséð	[Rare (classified poetic)]
afarvogun hvorugkynsnafnorð. Sjaldséð	[Rare (classified rare)]

The new version of the search results for *úkra** are shown in figure 3, marking the acceptability of the spelling variants with "Réttur ritháttur er ... " 'The correct spelling is ... ', cf. figure 1 in section 3 where the search heads are not shown. This immediately shows the users which spelling variants are standard and which are not.

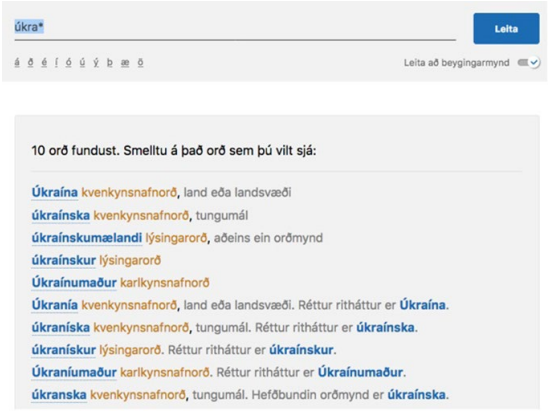


FIGURE 3. Searching for *úkra**, showing search heads.

5.2. Non-standard word forms and paradigms of errors

Neither the non-standard word forms nor the paradigms of errors are visible on the DMII website but they are to be used as supplementary data in the search to the benefit of the users. The goal is to make users aware of what is standard language and what is not, and the results from the non-standard word forms and paradigms of errors are only presented as references to the correct form. As of September 2022, this is still not visible for web users but will be added soon. The results will appear with the regular DMII results with the header "Þú gætir átt við ... " 'You might be looking for ... '.

Both headwords and inflectional forms are searchable in the DMII and search for inflectional forms returns a list of headwords containing that form. If a search string is found in both a regular DMII headword and as a non-standard word form or in a paradigm of errors, the results for the non-standard form are shown beneath the regular results. If the string is only found in the non-standard data, the results are beneath the standard message saying it was not found in the database. As an example, the form "alnar" is an inflectional form of the feminine noun *öln* 'ulna' but it is also a possible error form of four other words. In this case, the word *öln* would be listed as usual but after that, the other four headwords are listed as possibilities, with the caption 'You might be looking for':

öln kvenkynsnafnorð

Þú gætir átt við 'You might be looking for':
ala ól ólum alið, sagnorð
alin kvenkynsnafnorð
alinn lýsingarorð
álna álnaði álnað, sagnorð

The aim is to show the users that their search string is a standard form of the first word but only similar to (or an error form of) the other four. It is then up to the user to determine which headword they are actually looking for.

5.3. Unsolved problems

Questions of word boundaries are the reason for some of the most common types of real language errors in the search strings. These are difficult

to cope with in the DMII, which was originally strictly based on single-word paradigms.

The first type of error is splitting compounds in the search strings. This type of error is outside the scope of the DMII at present since all possible variations of erroneously split compounds cannot be added to the data. Using a compound splitter (Daðason et al. 2020) in "reverse mode" on the search strings might work, in the form of a suggestion, similar to how the non-standard word forms and paradigms of errors are presented. Examples of split compounds in the search strings are *lyfja afhending* for *lyfja-afhending* 'delivery of drugs', and *almennings stöðum* (dative), for *almenningsstöðum* 'public places'.

The other type is joining words erroneously, as in writing prepositional phrases and phrasal adverbials as continuous strings: *afhverju* (prepositional phrase, incorrect) for *af hverju* (correct form) 'why'; *einskonar* (adv., incorrect) for *eins konar* (correct form) 'some kind of'. Errors are also common in word class dependant word boundaries, as specified in the Spelling Rules: *ofstór* adj. (incorrect) for *of stór* (correct form) 'too big'. (The spelling rule specifies that the adverb *of* 'too' is concatenated to nouns and verbs, but a free form preceding adjectives and adverbs.)

The problem is that multiword headwords cannot be added in all possible cases, but work is in progress on finding a method of presenting suggestions based on current LT tools for Icelandic in order to give the users hints on the nature of their search string errors, in the form: "You might be looking for ..." or "The correct form might be ...".

6. Conclusion

This paper has described some of the steps taken recently in the development of the DMII from being purely descriptive towards being prescriptive. In the last few years, the focus of the project has been on LT, but it is now shifting to the users of the website, looking at what they really search for and how they may be misreading the information. The remedy proposed in this paper is to shift as much information as possible to the earliest possible place in the search process. In that manner, the search heads and added references from errors to standard forms in the search lists should help users looking to correct their grammar and spelling, even when they might have a tendency to jump to conclusions.

References

Almannarómur. <https://almannaromur.is/en>. Accessed September 2022.

Beygingarlýsing íslensks nútímamáls. [*The Database of Modern Icelandic Inflection*.] <bin.arnastofnun.is>. (no date) Kristín Bjarnadóttir, editor. The Árni Magnússon Institute for Icelandic Studies.

Bjarnadóttir, Kristín, & Hlynsdóttir, Kristín Ingibjörg. 2020. Online Data on Icelandic Inflection: Descriptive to Prescriptive: "Why, for whom, by whom" and how? *Nordiska studier i lexikografi* 15. Rapport från 15 konferensen om lexikografi i Norden. Helsingfors 4–7 juni 2019, pp. 71–79.

Daðason, Jón Friðrik, Mollberg, David Erik, Loftsson, Hrafn, Bjarnadóttir, Kristín. 2020. Kvistur 2.0: a BiLSTM Compound Splitter for Icelandic. *LREC 2020 Proceedings,* pp. 3984–3988.

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