DIM: The Database of Icelandic Morphology

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
University of Iceland
kristinb@hi.is, kih4@hi.is, steinst@hi.is

Abstract

The topic of this paper is The Database of Icelandic Morphology (DIM), a multipurpose linguistic resource, created for use in language technology, as a reference for the general public in Iceland, and for use in research on the Icelandic language. DIM contains inflectional paradigms and analysis of word formation, with a vocabulary of approx. 287,000 lemmas. DIM is based on The Database of Modern Icelandic Inflection, which has been in use since 2004. Whereas the older work was descriptive, the new version is partly prescriptive, making the data applicable in a greater range of projects than before.

1 Introduction

This paper describes The Database of Icelandic Morphology (DIM), containing the morphological analysis of approx. 287,000 Icelandic lemmas. The DIM is based on The Database of Modern Icelandic Inflection (DMII), a collection of inflectional paradigms first published in 2004, and originally conceived as a resource for language technology (LT) (Bjarnadóttir, 2012). The DMII has been restructured and extended to include information on word formation, and the analysis has been extended to include genre, style, domain, age, and various grammatical features. The original DMII was descriptive, but DIM is partly prescriptive, i.e., the "correctness" of both words and inflectional forms is marked in accordance with accepted rules of usage. This greatly improves the scope of applications using the data, from the purely analytical possibilities of the old DMII (used for e.g. search engines, PoS tagging, named entity recognition, etc.), to the productive possibilities of the DIM, such as correction and formulation of text. The additional analysis of morphological constituent structure also provides important linkups between lexical items, as the morphology of Icelandic is extremely productive. The name DIM is here used inclusively for the new project, whereas DMII refers to the inflectional part only.[1]

DIM has five aspects:

- An LT data source for various uses (inclusive of the original format available from 2007)
- A new enhanced and enlarged website for the general public
- The prescriptive DMII Core which is a subset of the inflectional paradigms marked for correctness
- A morphological analysis (MorphIce) with binary constituent structure and lemmatization of constituents
- A data source for linguistic research utilizing the classifications in the database to the full.

The paper is structured as follows. Section 2 contains a short description of Icelandic morphology, to pinpoint the features of analysis needed for various LT uses, with a discussion of the difference of descriptive and prescriptive data. Section 3 describes the DIM database and discusses details of the classification system briefly. Section 4 describes the five main parts of DIM listed above, one by one, drawing out the benefits of the new classification system in each case, i.e., in the DIM

[1] The Icelandic name of the DMII is Beygingarlýsing íslensks nútímamáls, abbreviated BÍN. The abbreviation has become a household name in Iceland, with the noun BÍN assigned feminine gender, and the verb bína (with the object in the accusative) used of the search. As the name is so well known, it is not easy to rename the project, and thus BÍN is still used inclusively in Icelandic for the whole project, both DMII and DIM, i.e., BÍN-vefurinn (DMII Web, for inflection online), BÍN-máltæknigögn (DIM/DMII LT Data), BÍN-kjarninn (DMII Core), BÍN-orðföng (DIM Morphological Data, MorphIce)), and Rannsóknar-BÍN (DIM for Research).

Core, the DMII Web, MorphIce, the accessible LT data, and a website for linguists doing research on Icelandic. Section 5 gives details on availability and licensing, and Section 6 contains the conclusion.

2 DIM and the Morphology of Icelandic

Work on the DMII was started in 2002, at the Institute of Lexicography in Reykjavík (cf. Bjarnadóttir (2012) for a description of the project).[2] The database (and the analysis) was very limited in scope, due to considerations of finance and manpower. The result was a set of inflectional paradigms for Icelandic intended for LT use, and a website for the general public.[3] The downloadable LT data has been available in two formats, i.e., in a list of inflectional forms with grammatical tags, linked to a list of lemmas, and in a simple list of inflectional forms without any analysis. Up to date, the data has most popularly been downloaded as a CSV file, with the fields lemma; word class; domain; inflectional form; grammatical tag. The inflectional data for the genitive singular definite form of the masculine noun **köttur** "cat" is as follows, slightly simplified, with the grammatical tag in English:

köttur;416784;masc;kattarins;GEN.SG.+DEF;

This simple data has been used extensively in Icelandic LT projects to date, but these projects have shown the need for a more extensive analysis of Icelandic morphology, which is rich and full of variants and ambiguities, both in regard to inflection and word formation. The reasons are shortly addressed in the following subsections, insofar as they are reflected in the analysis used in DIM.

2.1 Inflection

The ratio of inflectional forms to paradigms in the original DMII is quite high, i.e., 5.8 million inflected forms to 270,000 paradigms, with up to 16 inflectional forms to a noun, 120 to an adjective, and 107 to a verb, excluding variants (Bjarnadóttir, 2012).[4] The inflectional categories for nouns are case (nom., acc., dat., gen.), number (sg., pl.), and definiteness (-/+);[5] for adjectives gender (masc., fem., neut.), case (4), number (2), definiteness (2), and degree (pos., comp., superl.); for finite verbs voice (active, mediopassive), mood (indicative, subjunctive), tense (present, past), number (sg., pl.) and person (1st, 2nd, 3rd).[6] For other word classes, some adverbs inflect for degree; personal pronouns inflect for person, case and number; other pronouns, the definite article and the numbers from one to four inflect for gender, case and number, etc. All these features are used in the tag set for the DMII, which is correspondingly large, cf. footnote 4.

The number of inflectional forms is, per se, not problematic, but the number of variant forms with the same grammatical tag within the same paradigm can be. A simple case in point is genitive forms taking different endings, as in the genitive singular of the masculine noun **lestur** "reading": **lestrar/lesturs**. These two genitive forms are equally acceptable, but restrictions on the usage of variants can, in other lexical items, be a question of context, style, and degree of acceptability. A case in point is the feminine noun **rödd** "voice" where the otherwise obsolete dative singular variant **röddu** is only used in contexts like **hárri röddu** "with a forceful voice" (i.e., "loudly, clearly"; this instrumental dative phrase construction is quite common). The result of the number of possible variants of inflectional endings (i.e., exponents of a grammatical category) is a rather large number of inflectional patterns or inflectional classes.[7]

The inflectional forms are highly ambiguous, and this can be demonstrated by the distribution of the inflectional endings. In the genitive plu-

[2] In 2006, The Institute of Lexicography merged with other institutions under the name The Árni Magnússon Institute for Icelandic Studies.

[3] It should be noted that the DMII is a set of hardcoded paradigms and not a rule-based inflectional system. The reasons for this are given in Bjarnadóttir (2012).

[4] The figures are from Bjarnadóttir (2012), but the number of inflectional forms of verbs quoted here is from the original DMII. The new DIM includes additional cliticized verb forms, and grammatical tags for impersonal constructions (i.e., verbs with oblique subjects), so the number of possible inflectional forms for a verb presently exceeds 300. That analysis is under review, and there are plans to review the PoS tag set for Icelandic which at present contains more than 670 possible morphosyntactic tags, of which 559 turn up in a corpus of 1.2 billion running words (Steingrímsson et al., 2018).

[5] Gender is a lexical category for nouns, not an inflectional one.

[6] The categories for non-finite verbs are not listed here.

[7] The paradigm of each word is run as a whole, i.e., all the inflectional variants are produced by one bundle of rules, instead of specifying that a word can belong to more than one inflectional class. There are at present (May 2019) 669 such inflectional classes in the DIM; the number fluctuates easily with new data.

ral, the universal ending for nouns is **-a**[8] (as in **anda**, gen.pl. for both the feminine noun **önd** "duck" and the masculine noun **andi** "spirit"), but the same ending is also one of the nom.sg. endings in feminine nouns, the acc./dat./gen.sg. ending in some masculine nouns, the acc.pl. ending in some masculine nouns, not to mention the function of the same ending in other word classes. The result is that inflectional forms are hugely ambiguous, with only 32% of the inflectional forms in the original DMII being unambiguous (Bjarnadóttir, 2012). Disambiguation is therefore an important task in Icelandic LT, and because of the idiosyncrasies of individual words in respect of variant inflectional forms this can only be achieved by referring to a lexicon.

2.2 Word Formation

In Icelandic, the morphological head of a word is the word-final base word or compound, depending on the binary structure of the word (cf. Bjarnadóttir (2017a) for a short reference). Compounds can be formed by joining any of the open word classes, but noun-noun compounds are by far the most common. The rules of compound formation are recursive, and there is no theoretical limit to the number of constituents in a compound, although compounds with more than six constituents are rare. An added complication is the fact that the first part of compounds (i.e., the modifier) can take a variety of combining forms. Nominal modifiers can appear as stems or inflected forms, most often in the genitive, singular or plural. The choice of forms is arbitrary, but not free, cf. examples in Table 1 where unacceptable compounds are marked by * (cf. also Bjarnadóttir (1995)).

Stem	Gen.sg.	Gen.pl.	Meaning
bóksala	*bókarsala	*bókasala	"book store"
*bókkápa	bókarkápa	bókakápa	"book cover"
*bókbúð	*bókarbúð	bókabúð	"book store"

Table 1: Examples of combining forms in Icelandic compounds.

The lemmatization of the modifiers is needed for disambiguation, as inflectional forms are highly ambiguous, as in the case of the genitive plural **anda** in **andagift** "spiritual gift" (i.e., "inspiration"), and **andapollur** "duck pond", where the lemma for **anda** in the first compound is the masculine noun **andi** "spirit, breath", but the feminine noun **önd** "duck" in the second compound.

The ambiguity of the combining forms is a reflection of the ambiguity of inflectional forms, and the most ambiguous of those in the DMII at present is **minni**, which shows up as 30 inflectional forms in four paradigms, i.e., in the neuter noun **minni** "memory" (5 inflectional forms: nom./acc./dat. sg., nom./acc.pl.); in the verb **minna** "remember" (4 inflectional forms, not counting impersonal ones: active voice, 1.p.sg. pres. indicative & subjunctive; 3.p.pl. pres. subjunctive); in the adjective **lítill** "small, little" (20 inflectional forms, i.e., comp., masc./fem. nom./acc./dat./gen.sg., and masc./fem./neut. nom./acc./dat./gen.pl.); and in the possessive pronoun **minn** "mine" (1 inflectional form, fem.dat.sg.). The ambiguity in combining forms is therefore linked to ambiguity elsewhere in the DIM.

2.3 Description vs. prescription

The DMII was originally created as a part of an effort to start work on LT at the start of the millennium, financed by the Icelandic Ministry for Education and Culture. The first version of the DMII was a set of XML files with 173,389 paradigms, made available on CDs for use in LT in 2004. The purpose was quite simple, the data was to be used in coping with the morphology in Icelandic texts, as well as in search engines and other tools requiring the information. The emphasis was being able to cope with Icelandic texts "as is", without regard to correct spelling, grammar, vocabulary, or style. In other words, the data was descriptive, as it had to function for analysis, but it was in no way suited for production. For that prescriptive data is needed, in order to conform with the established standards for good Icelandic.

Icelandic standards appear in the "Rules of spelling and punctuation" (*Ritreglur* (2016), *Reglur um greinamerkjasetningu* (2018)), published by the Ministry of Education and Culture, and in *Stafsetningarorðabókin* "The Dictionary of Spelling" (Sigtryggsson, 2016), published by the Árni Magnússon Institute for Icelandic Studies.[9] Various handbooks and grammar books, used in

[8] Except for a subset of feminine nouns and a (very) few neuter nouns where the generative plural ending is **-na**.

[9] These sources are available at the website of The Árni Magnússon Institute for Icelandic Studies, `https://arnastofnun.is` and at `https://malid.is`.

the school system, also function as prescriptive sources.

In order to make the transition from descriptive to prescriptive, the core vocabulary of the DMII has been checked against the standards mentioned above, adding extensive cross-referencing of less than optimal instances of usage to the standardized forms in the DMII. This applies both to spelling, inflectional forms, and to the vocabulary itself, as a part of good usage is considered to be in the choice of native words instead of loan words.

The mark-up of the prescriptive data, with the links from the less optimal forms to the prescribed ones, makes the data better suited for LT projects such as spell checkers, grammar checkers, and any kind of production of text, including teaching material such as grammatical exercises.

3 The structure of the DIM Database

The old DMII database has been restructured to achieve two main objectives. The first is to allow for more detailed information and classification (cf. Subsection 3.1), and the other to allow for configuring different aspects of the data for different use (cf. Section 4). Some parts of the new database have been designed from scratch and contain new types of data, such as the analysis of word formation which is linked to the other parts of DIM.

Examples of the detailed information added to the database are features of usage restrictions (syntactic, semantic, stylistic, etc.), different orthographic representations, and features of acceptability, as described in Subsection 3.1. In addition, a proposed part of the DIM, still only in the preliminary stages, is a repository of written word forms not found elsewhere in the DIM data, including obsolete forms, errors of all kinds, and abbreviations. These will be classified, dated, attributed to source, and be linked to the list of lemmas in the DIM proper. Extensive material of this type is ready for import into the database, both from lexicographic sources and from error analysis. This data allows for new analytic possibilities, both for language technology and linguistic research. It also extends the time frame of the data, by creating a place for older inflectional variants, as the DMII was originally confined to Modern Icelandic, cf. the name: The Database of Modern Icelandic Inflection. The inclusion of older data does not entail attempts at creating paradigms for Icelandic through the centuries; the data is too scarce for that. Experiments with using these peripheral word forms in LT have been made, cf. footnote 14 below.

3.1 Classification

DIM uses a new sorting and grading system for words and inflectional forms to differentiate between prescriptive and descriptive use, or to give researchers access to vocabulary containing relevant grammatical features. Following is a brief description of the main sorting categories, with a handful of examples in footnotes. The system is complex and a full exposition with examples is outside the scope here.

- **Grammatical features of words**: Used to mark words with certain features or restricted usage. Words can be marked for more than one feature. These include: Idiom bound, gender variation, older word form, restricted paradigm, loan word, spelling variants ...[10]

- **Value of inflectional forms**: Used for inflectional forms where two or more variants are presented, to indicate their status in respect of the other variant(s), with values like: equal, dominant, yielding, uncertain.[11]

- **Correctness Grade**: Used to mark a word's or variant's correctness according to prescriptive grammar rules and standardized spelling. Grades range from 0 to 5. Most words have a grade of 1, and this is the default value and stands for "Correct". Grades 2, 3 and 4 stand for "Used", "Not good" and "Very bad", depending on the level of "wrongness", 4 being the lowest grade. Grade 5 exists for words or inflectional forms that have somehow made it into the database but are so "wrong" that they

[10] Examples of "Grammatical features of words": Idiom bound: **almannavitorð** n.neut. "public knowledge", only used in the dative singular in the idiom *e-ð er á almannavitorði* "sth is public knowledge"; Gender variation: **engifer** n.masc. or n.neut. "ginger"; Older word form: **röddu** dat.sg. of **rödd** n.fem. "voice", cf. Subsection 2.1; Restricted paradigm: **munu** v. auxilary. The verb is finite only, active voice only, and there is no past tense in the indicative; Loan word: **engifer** n.masc. or n.neut. "ginger" (Loanwords, especially multisyllable ones, need marking as their inflection very often deviates from the inflection of the native vocabulary.); Spelling variants: **pósitívur/pósitívur** adj. "positive".

[11] Examples of "Values of inflectional forms": Equal: **dugir/dugar**, 3.p.sg.pres. active voice of **duga** v. "suffice"; Dominant: **rödd**, dat.sg. of **rödd** n.fem. "voice"; Yielding: **röddu**, dat.sg. (idiom bound) of **rödd** n.fem. "voice".

could not possibly be part of the language of an adult with a native speaker's competence in Icelandic, and so they are only visible in administrator mode. Words and inflectional variants classified by genre (cf. next paragraph) have the grade 0, standing for "Not applicable, depends on style or genre". This is because variants marked by genre are not incorrect, but they are not the most common correct form either.[12]

- **Genre**: Used to sort words and inflectional forms according to style or age. Values: Formal, informal, derogatory, obscene, rare, old-fashioned, obsolete, poetic language, regional. Genre is not a mandatory feature for words or variants.

- **Domain**: A semantic classification, used to classify named entities and domain specific vocabulary. These include several different kinds of names (e.g. Icelandic personal names, place names, etc.), and technical terms from different fields, etc. Most words are in the domain called Common language, which is the default value. All words belong to one domain only.[13] Domain specification only applies to words, not to inflectional forms.

- **Pronunciation**: Features of pronunciation are marked on words with possible discrepancies between pronunciation and spelling. Still a work in progress, this is meant for linguistic research, but it may be useful for speech synthesizers or speech analysis, etc.

- **Peripheral word forms:** Older word forms, spelling errors and other forms that do not fit within a paradigm are kept in a separate list and connected to a lemma in the database. This can be used to connect errors or old forms to the appropriate modern form.[14]

[12]Examples of "Correctness Grade": 1 (Correct, default value): **jafnvægi** n.neut. "balance". 2 (Used): **ballans** n.masc. "balance". The compound **jafnvægi** is considered better usage, it is also much more common.; 3 (Not good): **pósitífur** adj. "positive". According to *The Dictionary of Spelling* the correct form is **pósitívur**; 4 (Very bad): **líter** n.masc. "liter". The correct form is **lítri** n.masc. 0 (Grading is not applicable): **blóðgagl** "raven, vulture", poetic language (**blóð** "blood" and **gagl** "goose; bird").

[13]The result is that the common noun **hrafn** "raven" and the personal name **Hrafn** are shown in two different paradigms, although the inflection is identical.

4 The Five Aspects of DIM

DIM has five aspects or conceptual units. All of them are part of one database; two of them are solely focused on inflection, i.e., the DMII Core (Subsection 4.1) and the DMII Website (Subsection 4.2), and one contains the word formations analysis, i.e., MorphIce (Subsection 4.3). The remaining two aspects contain different modes of access to the data, the LT Website for use in language technology (Subsection 4.4), and the LingResearch for research on the language (Subsection 4.5).

4.1 The DMII Core

The DMII Core is designed to meet users' demands for data from the DMII in a prescriptive context, and it is created to be used for third party publication through an API, especially for language learners. The RESTful API is open for everyone to use. It allows users to send simple queries and receive full paradigms in JSON-format as a response. The data in the DMII Core only contains the core vocabulary of Icelandic, only standardized spelling, and only the most common or correct inflectional forms. This makes the data suited for creating teaching material and for other prescriptive uses. As the data in the DMII Core is simplified and the vocabulary limited, the omission of a word or variant does not imply that it is wrong. However, if a word or variant is found in the DMII Core, users should be able to trust that it is safe to use, i.e., correct, in all (or most) contexts.

The vocabulary of the DMII Core is based on the list of headwords in *The Modern Icelandic Dictionary* (Jónsdóttir and Úlfarsdóttir, 2016), and the 50,000 most frequent words (lemmas) in the *The Icelandic Gigaword Corpus* (Steingrímsson et al., 2018). The total number of paradigms in the DMII Core now stands at 56,867 (end of May 2019), as compared almost 287,000 in the whole of DIM.

The new classification system (cf. Subsection 3.1), is used to choose words and inflectional variants for the DMII Core. Only words and variants with a correctness grade of 1 (universally acceptable) are included in the DMII Core,

[14]Cf. Daðason et al. (2014) for description of an LT tool for transposing older Icelandic texts to modern spelling, using data from a pilot project of this kind.

and the categories of genre included are "formal, informal, derogatory, obscene". (The excluded ones are "rare, old-fashioned, obsolete, poetic language, regional".) The only domain included in the DMII Core is common language (i.e., the default value), with a chosen selection of named entities, i.e., common Icelandic personal names, a few very common place names, and the most common names of institutions.

The paradigms in the DMII Core have been simplified as possible, without omitting equally valid variants, showing only the best forms, or the variants not limited by specific usage restrictions. The correctness grade is also used for inflectional forms, and only variants with correctness grade of 1 are included.

4.2 The DMII website

Individual paradigms have been accessible on the DMII website from 2004.[15] Extensively used by the Icelandic public as a reference on inflection, the website has been popular from the start, and the latest figures show that more than 200,000 users viewed over 1.7 million pages in the year starting June 1, 2018. (The total population of Iceland is approx. 360,000.) The figures are still rising, with 9% growth over the previous year. Originally, the data was set out to be purely descriptive, for use in analysing Icelandic text "as is", not just the "received" text adhering to the rather strict language norms officially advocated. To make the website more useful, notes on usage were placed with individual inflectional paradigms, pointing the way in the choice of variants, and containing information on restrictions on their use. These notes are in Icelandic only, and they were originally hand-crafted and not classified in any way. This makes the original website unsuited for any but speakers with native or near native knowledge of Icelandic, as the users themselves have to make the final choice between variants, with the help of the notes. In this context, the multitude and ambiguity of variant forms has to be stressed.

This mode of operation has not always been totally successful, as the descriptive nature of the data causes problems, even for native speakers of Icelandic. The expectation is that all word forms appearing on the site are "best usage", all spelling variants are "good", and all words shown on the website are acceptable, irrespective of genre, style, etc. The tolerance for substandard usage appearing on the website is at times very low, as the users will let the editors know from time to time. The converse is also true, as some users expect all "acceptable" Icelandic words to be found in the DMII, in spite of clear statements to the contrary on the website, stating that the DMII is not exhaustive and not an authority on acceptable Icelandic vocabulary. The original duality of purpose in the DMII, i.e., the need for the maximum number of word forms found in texts for LT analysis vs. the needs of ordinary users for prescriptive data, has therefore caused some problems from the start.

The new version of the DMII website, to be opened in 2019, makes use of the work on classification in the DMII Core described in Subsection 3.1. Markings on variations in spelling and word formation, and any restrictions on use discovered in the work of adapting the standard, are carried over into the paradigms on the DMII website, in the form of better notes on usage, with extensive cross-references. To give one example of two words commonly causing confusion because of their spelling, the words **híði** "den" (as in "bear's den") and **hýði** "skin" (as in "banana skin") are pronounced in the same way. Each of these entries on the website gives both forms, explaining the semantic difference. In case of substandard spelling (as in writing **pósitífur** instead of **pósitívur** "positive"), a hyperlink to the appropriate spelling rule is provided, as published online at The Árni Magnússon Institute's portal for information on Icelandic usage.[16] The guidelines on the DMII website still warn the users that the DMII is **not** a spelling dictionary, but there are now referrals to the standard wherever possible.

4.3 MorphIce

The morphological analysis included in MorphIce gives full constituent structure, with lemmatized constituents. As compounding in Icelandic is extremely productive, this is of importance in LT tasks, as the data can be used to minimize the effect of out-of-vocabulary words. The data also serves for training of tools such as the compound splitter Kvistur (cf. Daðason and Bjarnadóttir (2014)), which has been used to estimate the probability of unknown compounds by the use of a preliminary version of the data in MorphIce. This data was originally analysed manually in the

[15] http://bin.arnastofnun.is

[16] https://malid.is

nineties (Bjarnadóttir, 2006) and will now be incorporated into the DIM. It is only with the creation of MorphIce that this data can be made freely available as a linguistic resource.

As stated in Subsection 2.2, compound formation in Icelandic is complex. The compound analysis in MorphIce assumes binary branching, and the rules are recursive, resulting in quite complex binary trees.

The data on each compound will give full details on each component, with links to the main index in the DIM, with DMII ids. All information on the inflection of each constituent word in the compound will therefore be accessible.[17] Bound constituents (affixes and combining forms) form a separate part of the data set, with information on the structures into which they fit. The format of the output is still under construction, but a sample of the analysis can be seen in the word **orðabókarmaður** "lexicographer" below, with the consituent parts **orð** "word", **bók** "book", and **maður** "man". The analysis is binary, with each item in the example showing the result of one process or word formation rule. These can be nested or not in the output, according to needs, but as stated above, the final format is still under construction. The numbers are the ids of the words in the DMII.

- orðabókarmaður:
  [[orðabókar]<gen.sg.orðabók.n.fem.404616>
  [maður]<n.masc.5763>]<n.masc.88516>

- orðabók:
  [[orða]<gen.pl.orð.n.neut.2635>
  [bók]<n.fem.11100>]<n.fem.404616>

A coding system for the binary trees is included in the data, with "0" for base words, "1" for a single join, "12" for a left-branching binary tree with three constituents, as in [[[orða][bókar]][maður]] above, etc. The granularity of the compound analysis can by adapted to the needs of each LT task, and the most detailed analysis will probably only be of interest to linguists.

The data in MorphIce will be linked to a dataset containing argument structure, presently found in a pilot version on the website of The Árni Magnússon Institute for Icelandic Studies (Bjarnadóttir, 2017b). This data will be used to link multiword constructions, such as particle verbs and verbs with incorporated object, to their compound counterparts, as in **greina að** "separate" (i.e., "take apart", the compound verb **aðgreina** also exists), the past participle/adjective **aðgreindur**, the present participle **aðgreinandi**, and the noun **aðgreining**. Structures of this kind are very common in Icelandic, and finding and analysing multiword lexical entities and linking them to compounds helps in demarcating semantic units.

In the future, the plan is to make the binary trees themselves accessible online, but as yet the formulation of that project is only in the preliminary stages.

4.4 The LT Website

Datasets from DIM, for use in language technology, are available from a separate website. Until now, two versions of the data have been available for download on the old DMII website along with detailed descriptions of the datasets, in Icelandic and English. One version is the list of inflectional forms and lemmas, along with word class and grammatical tag, described in Section 2. The other is a simple list of word forms, without any classification or linkups. The new DIM-LT website still makes the DIM data available for download in those formats, but more configurations are available, constructed in cooperation with a select group of long-time users. The data available on the DIM-LT website is updated daily. For reproduction purposes, versioned datasets will be published periodically.

The dataset is published with an open license and is intended for use in the development of LT tools and methods, but it may also be suitable in other fields as well. The DIM-LT data is not intended for lookups. Users who build software that needs to do lookups in DIM at runtime will be encouraged to use the API, described in Sections 4.1 and 5, as that will allow them access at all times to the most recently updated data.

As well as providing access to downloadable data, the DIM-LT website has information on licensing, detailed information on all the grammatical features appearing in the inflectional paradigms, including lists of word classes, all inflectional categories and grammatical terms, and

[17] The inflection of Icelandic compounds is notoriously unpredictable, especially when there is a choice of variant forms, as in **útvegur** "fishing industry", gen.sg. **útvegs** (i.e., **út** "out" adv., **vegur** "road, way") vs. **akvegur** "road (for cars)", gen.sg. **akvegar** (**aka** "drive" verb, **vegur** "road"). In recursive compounding, such genitives do appear as modifiers. In such cases, the data should be sufficient for analysis, but perhaps not unerring in predictions for new compounds.

other relevant information, in Icelandic and English. The tag set in the downloadable data is in Icelandic, but English translations are accessible on the website.

4.5 DIM LingResearch

The DIM LingResearch is an adaptation of the editorial interface to the new database, to be made accessible for linguistic research. All features of the classification of words and inflectional forms will be accessible for the extraction of data. Some features have been specially included in the database with linguistic research in mind, such as categories of irregular pronunciation. To name an example, the vowel **a** is diphthongized when preceding **ng** (as in the word **bangsi** "teddy bear"), except in a few loanwords, such as **mangó** and **tangó**. This type of data is interesting for linguists working on morphophonemics, but it may also be useful for speech systems, etc. There are at present 16 pronunciation sets of this kind in the data, along with similar sets for various other features of word formation, etc.

5 Access and licencing

As described in previous sections, DIM is available in different configurations. The DMII Core is made available through an API through a permissive license that allows third parties to publish the data on the web. They are required not to modify the data and to give appropriate credit. All paradigms can be viewed individually on the new DMII website, as they have been on the old DMII website since 2004. Access is open to all and free of charge, but scraping the data or copying en masse is prohibited. For use in language technology or research on the language, the datasets are available for download with an open, permissive license, CC BY-SA. The same will apply to the MorphIce data. Finally, access to the DIM LingResearch interface, intended for scholars, will be given upon request.

6 Conclusion

The data from the old DMII has been used extensively from 2004, when it was first made available. The initial dream was basically to create data for rather simple tools, like a decent search engine, etc. Very many of the projects using the data have far surpassed these expectations. Two of these projects are mentioned above, i.e., the compound splitter Kvistur (cf. Subsection 4.3), and the spellchecker Skrambi, referred to in connection with the use of peripheral word forms (cf. Subsection 3.1, footnote 14).[18] Two very recent NLP-tools take advantage of DMII, the PoS Tagger ABLTagger (Steingrímsson et al., 2019) and the lemmatizer Nefnir, which is described in this current version of NoDaLiDa (Ingólfsdóttir et al., 2019). A list of additional projects is accessible on the DIM website.[19]

In fall 2019, the first projects will start in a national language technology programme, which will run for five years. The programme follows a plan set forward in 2017 (Nikulásdóttir et al., 2017), and aims to produce open systems for machine translation, spell and grammar checking, speech synthesis and speech recognition.

Extensive linguistic resources are needed for the new projects planned in the next five years. This is especially important for two reasons. First, the Icelandic language community is very small, and although Icelanders take pride in the production of a great deal of text (as evidenced in a blooming Icelandic literary scene, and the proliferation of Icelandic websites, etc.), the actual mass of text produced is nowhere close to the scale accessible in really large language communities. Even if all Icelandic texts from all times were accessible, very many word forms would probably only occur a few times, certainly not often enough to be useful in statistical analysis.[20] The other reason is the complexity of Icelandic morphology, both inflectional and morphological, with corresponding irregularities and ambiguities. It has therefore proved to be necessary to produce and store the morphological data, instead of writing a rule system for analysing the morphology on the go.

We see the work in the immediate future as ongoing excerption from the Gigaword Corpus (Steingrímsson et al., 2018), fine-tuning the analysis and classifications described in this paper, and ongoing cooperation with the users of the data, as they are the people who know what they need.

[18] Skrambi is available as an online spellchecker (`http:skrambi.bin.arnastofnun.is`), but it is also used in different versions for tasks such as correcting OCR texts, and to transpose older texts with unstandardized spelling from different periods to Modern Icelandic.

[19] `http:bin.arnastofnun.is`

[20] Some inflectional forms are not found at all, as is the case for the dative singular of the name of Odin's tree **Yggdrasill** which is not to be found in any of the Old Icelandic sources.

Acknowledgments

The restructuring of the DMII and the creation of DIM was made possible by a grant by the Icelandic Research Council in 2017. The authors wish to thank our co-workers in the project, without whom the project would not have taken shape: Samúel Þórisson, for the database, and Trausti Dagsson, for the creation of the new websites. Thanks are also due to our present and former colleagues at The Árni Magnússon Institute for Icelandic Studies and The University of Iceland who have been generous with their time, advice and assistance: Starkaður Barkarson, Jón Friðrik Daðason, Sigrún Helgadóttir, Halldóra Jónsdóttir, Jón Hilmar Jónsson, Ari Páll Kristinsson, Kristján Rúnarsson, Eiríkur Rögnvaldsson, Jóhannes B. Sigtryggsson, Einar Freyr Sigurðsson, Ásta Svavarsdóttir, Þórdís Úlfarsdóttir, Ágústa Þorbergsdóttir, Gunnar Thor Örnólfsson, and Katrín Axelsdóttir. Thanks are also due to the company Já.is for their generous support through the years.

References

Kristín Bjarnadóttir. 1995. Lexicalization and the Selection of Compounds for a Bilingual Icelandic Dictionary Base. *Nordiske studier i leksikografi*, (3):255–263.

Kristín Bjarnadóttir. 2006. *Afleiðsla og samsetning í generatífri málfræði og greining á íslenskum gögnum*. Orðabók Háskólans, Reykjavík, Iceland.

Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. In *LREC 2012 Proceedings: Proceedings of "Language Technology for Normalization of Less-Resourced Languages", SaLTMiL 8 – AfLaT*, pages 67–72.

Kristín Bjarnadóttir. 2017a. Phrasal compounds in Modern Icelandic with reference to Icelandic word formation in general. In Carola Trips and Jaklin Kornfilt, editors, *Further investigations into the nature of phrasal compounding*. Language Science Press, Berlin, Germany.

Kristín Bjarnadóttir. 2017b. https://notendur.hi.is/ kristinb/divs-2017.txt *ÍSLEX-venslamálfræði*. The Árni Magnússon Institute for Icelandic Studies, Reykjavík, Iceland.

Jón Friðrik Daðason and Kristín Bjarnadóttir. 2014. Utilizing constituent structure for compound analysis. In *Proceedings of LREC 2014. Reykjavik: ELRA*, pages 1637–1641.

Jón Friðrik Daðason, Kristín Bjarnadóttir, and Kristján Rúnarsson. 2014. The Journal Fjölnir for everyone: The post-processing of historical OCR texts. In *Proceedings of Language Resources and Technologies for Processing and Linking Historical Documents and Archives - Deploying Linked Open Data in Cultural Heritage. Reykjavik: ELRA*, pages 56–62.

Svanhvít Ingólfsdóttir, Hrafn Loftsson, Jón Daðason, and Kristín Bjarnadóttir. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the* 22^{nd} *Nordic Conference of Computational Linguistics*, NODALIDA 2019, Turku, Finland.

Halldóra Jónsdóttir and Þórdís Úlfarsdóttir, editors. 2016. http://islenskordabok.arnastofnun.is *Íslensk nútímamálsorðabók*. The Árni Magnússon Institute for Icelandic Studies.

Anna Björk Nikulásdóttir, Jón Guðnason, and Steinþór Steingrímsson. 2017. https://notendur.hi.is/eirikur/mlt-en.pdf *Language Technology for Icelandic 2018-2022: Project Plan*. Mennta- og menningarmálaráðuneytið, Reykjavík, Iceland.

Jóhannes B. Sigtryggsson, editor. 2016. http://malid.is *Stafsetningarorðabókin*. The Árni Magnússon Institute for Icelandic Studies.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of RANLP 2019*, Varna, Bulgaria.