The Database of Modern Icelandic Inflection (Beygingarlýsing íslensks nútímamáls)

Kristín Bjarnadóttir

The Árni Magnússon Institute for Icelandic Studies
Iceland
kristinb@hi.is

Abstract

The topic of this paper is the Database of Modern Icelandic Inflection (DMII), containing about 270,000 paradigms from Modern Icelandic, with over 5.8 million inflectional forms. The DMII was created as a multipurpose resource, for use in language technology, lexicography, and as an online resource for the general public. Icelandic is a morphologically rich language with a complex inflectional system, commonly exhibiting idiosyncratic inflectional variants. In spite of a long history of morphological research, none of the available sources had the necessary information for the making of a comprehensive and productive rule-based system with the coverage needed. Thus, the DMII was created as a database of paradigms showing all and only the inflectional variants of each word. The initial data used for the project was mostly lexicographic. The creation of a 25 million token corpus of Icelandic, the MÍM Corpus, has made it possible to use empirical data in the development of the DMII, resulting in extensive additions to the vocabulary. The data scarcity in the corpus, due to the enormous number of possible inflectional forms, proves how important it is to use both lexicographic data and a corpus to complement each other in an undertaking such as the DMII.

Keywords: Morphology, Inflectional database, Icelandic

1. Introduction

This paper describes the Database of Modern Icelandic Inflection (DMII; Beygingarlýsing íslensks nútímamáls), a collection of (at present) 270,000 paradigms with about 5.8 million inflectional forms, i.e., word forms with grammatical tags.[1] The DMII was initially created to serve two purposes, i.e., to produce data for use in LT projects, and to make the resulting paradigms available to the general public on the website of The Árni Magnússon Institute for Icelandic Studies (AMI).[2] From the outset the aim was to present Icelandic inflection ‘as is’, with as full a description of variants as possible. With this in mind, the decision was made to produce a full paradigm for as large a proportion of the vocabulary as possible, instead of producing a rule system for the generation of inflection by inflectional classes. It turned out that in spite of centuries of research on Icelandic morphology, the necessary data for a productive rule system was simply not available. The problem is that for analysis it may be acceptable to use an overgenerating rule-system, but for production it is not, if the end result, i.e., a text, is expected to be correct. This is a very relevant point, as demonstrated by the facts that the data from the DMII is used for context sensitive grammar correction, and the paradigms are also widely used online by the general public for reference. Native speakers of Icelandic need guidance to cope with a very complex inflectional system.[3]

Work on the DMII started in 2002, as a part of an LT Program launched by the Minister of Education, Science and Culture (Rögnvaldsson et al., 2009). The first version of the data was made available for LT use in 2004, and the online version was opened the same year. Data from the DMII has been used in various LT projects, such as search engines, PoS tagging, context sensitive correction, in language teaching, lexicography, etc. Both the DMII and the Tagged Icelandic Corpus (The MÍM Corpus) (Helgadóttir et al., 2012) are being produced at the AMI, and the two projects run in tandem. The original sources for the DMII were lexicographic, i.e., the electronic version of the classic *Dictionary of Icelandic* (Árnason, 2000), containing 135,000 headwords, and the AMI’s lexicographic archives. Various other sources are now used, but the next stage is to include the vocabulary contained in the MÍM Corpus. This work is now in progress.

The paper is structured as follows. Section 2 contains a short description of the richness of Icelandic morphology, followed in Section 3 by an account of the method used in creating the DMII and the two accessible versions of it, one for LT purposes and one online, for the general public. Section 4 contains an account of the limitation of the sources of information on Icelandic inflection, followed by Section 5, a description of the independent research needed to fill gaps in the sources. The inclusion of the vocabulary from the MÍM Corpus is described in Section 6, with the lesson learned on data scarcity in a language with a very rich morphology in Section 7. The conclusion is in Section 8.

2. The richness of Icelandic morphology

As can be inferred from the ratio of inflectional forms to paradigms in the DMII, i.e., 5.8 million inflected forms in 270,000 paradigms, the inflectional system of Icelandic is rich, with up to 16 inflectional forms to a noun, 120 to an

[1] The term Modern Icelandic is here used of contemporary Icelandic, i.e., 21st century usage.

[2] `http://bin.arnastofnun.is/`

[3] In November 2011 there were 268,011 pagewiews, and 52,570 visits from 60 countries to the online DMII. Iceland has about 320,000 inhabitants and most of the visits are domestic. The users are also in contact via email, with queries, additions and corrections.

adjective, and 107 to a verb, not including variants. This is reflected by the size of the tagset used in the PoS tagging of Icelandic, with over 700 tags (Pind et al., 1991) and (Helgadóttir et al., 2012).

The Icelandic inflectional system is also quite complex, as the endings that mark grammatical categories can, in some instances, have a number of variants, e.g., -s/-ar/-ur in the genitive singular of masculine nouns with a certain structure of base form, i.e., the ending -ur in the nominative singular. The result is a proliferation of inflectional variants, e.g., *þröskuldar/þröskulds*, genitive singular of the masculine noun *þröskuldur* 'threshold'. Furthermore, stem changes are common, both in vowels and consonants.

In the case of inflectional variants, the grammatical tradition in Iceland is to say that a word can belong to more than one inflectional class. However, the method of producing the paradigms for the DMII does not allow that; each lemma is shown in full in one paradigm, including all variants.[4] An inflectional class arrived at this way is in fact a unique bundle of inflectional rules, specific to a word or group of words.

3. The production of the paradigms

Initially, the paradigms in the DMII were produced with simple Unix shell scripts, by merging a matrix of inflectional endings containing slots for numbered variants of stems with records for individual words. The result was a set of XML files. The concept of the database now in use is similar. The record for each word contains all variants of the stem, and information on which parts of the full paradigm are applicable in each case ("flags"), e.g., no singular for pluralia tantum, no active voice in mediopassive verbs, no past participles for some verbs, etc. These records are merged with a matrix for the appropriate inflectional class and the resulting inflectional forms are then stored with morphosyntactic tags according to their place in the matrix.[5]

	Indefinite (+)			
	Singular (+)		Plural (+)	
Nom.	1+0	*akur*	3+ar	*akrar*
Acc.	1+0	*akur*	3+a	*akra*
Dat.	3+i	*akri*	2+um	*ökrum*
Gen.	1+s	*akurs*	4+a	*akra*
	Definite (+)			
Nom.	1+inn	*akurinn*	3+arnir	*akrarnir*
Acc.	1+inn	*akurinn*	3+ana	*akrana*
Dat.	3+inum	*akrinum*	2+unum	*ökrunum*
Gen.	1+sins	*akursins*	4+anna	*akranna*

Table 1: Matrix for the noun *akur*.

Table 1 shows a matrix for one class of nouns, with the resulting inflectional forms in italics. The word *akur* 'field, meadow' is flagged for the grammatical categories number and definiteness, i.e., +sg., +pl., +indef., +def., which specifies that there are no gaps in the paradigm. (For pluralia tantum (flagged −sg.), the singular would be left blank.) The table contains English translations of the Icelandic abbreviations used in the online version, which is similar to Table 1, leaving out the columns 2 and 4 (numbers for stems and the endings), retaining abbreviations and inflectional forms. The metalanguage is Icelandic.

The most commonly used output for LT purposes is a simple list with 6 fields, as in the 16 inflectional forms for the word *akur* 'field, meadow' in Table 2. The fields are lemma, identifier (number), word class or gender of nouns, type (i.e., common language, named entity, terminology, etc.; 'com' in Table 2 signifies 'common language'), inflectional form, and tag. The tags shown here are English translations, as in Table 1.

```
akur;472164;masc;com;akur;NOM-SG
akur;472164;masc;com;akurinn;NOM-SG-DEF
akur;472164;masc;com;akur;ACC-SG
akur;472164;masc;com;akurinn;ACC-SG-DEF
akur;472164;masc;com;akri;DAT-SG
akur;472164;masc;com;akrinum;DAT-SG-DEF
akur;472164;masc;com;akurs;GEN-SG
akur;472164;masc;com;akursins;GEN-SG-DEF
akur;472164;masc;com;akrar;NOM-PL
akur;472164;masc;com;akrarnir;NOM-PL-DEF
akur;472164;masc;com;akra;ACC-PL
akur;472164;masc;com;akrana;ACC-PL-DEF
akur;472164;masc;com;ökrum;DAT-PL
akur;472164;masc;com;ökrunum;DAT-PL-DEF
akur;472164;masc;com;akra;GEN-PL
akur;472164;masc;com;akranna;GEN-PL-DEF
```

Table 2: Output for LT: Example from a CSV file.

The matrices were (and still are) produced at need, every time a new variant makes a new inflectional pattern necessary, thus creating a new unique bundle of inflectional rules. This makes the description truly 'bottom-up', as it is purely based on the actual inflection of individual words.

There are at present over 630 such inflectional classes in the DMII, some with tens of thousands of words, but others showing the idiosyncracy of individual words, sometimes due to historical remnants of obsolete inflectional classes still attested in common phrases and idioms in the modern language.[6]

Out of the inflectional classes for nouns, adjectives and verbs, 42% contain no variants or other complicating features, such as internal inflection or unsystematic gapping.[7]

[4] The base form of the lemma is decisive in the division of paradigms. A variant base form will therefore produce two paradigms, as in the nouns *sannleikur/sannleiki* 'truth'. The base form for nouns is the nominative singular.

[5] For each word, intuition is used to choose between possible inflectional variants (i.e., between inflectional classes) and to assign values to the flags deciding the structure of the paradigm, e.g., +/−plural for nouns, +/−degree for adjectives, and +/−mediopassive for verbs, etc., when data is not available.

[6] The inflectional system of Icelandic has undergone some changes through the centuries, both structural changes (i.e., changes of inflectional classes per se), and drift of vocabulary between inflectional classes. As a whole, these changes are fairly minor, but the result is nevertheless very apparent in individual word forms.

[7] The DMII contains 49,296 pairs of variants and 262 triplets (Feb. 2012).

Lemmas	Infl. classes	Lemmas	Infl. classes
<10.000	5		
5.000-9.999	7		
1.000-4.999	29	<1000	41
500-999	21	<500	62
100-499	72	<100	134
50-99	47	<50	181
10-49	126	<10	307
2-9	146	<2	453
1	156		

Table 3: The productivity of inflectional classes (nouns, adjectives and verbs, 609 classes).

The number of words in each inflectional class varies greatly, as shown in Table 3, with about 25% of the inflectional classes of nouns, adjectives and verbs showing bundles of rules describing the truly idiosyncratic inflection of single words.

4. The source material and the lack of information

At the outset, the bulk of the source material for the project was from lexicographic resources, such as data from the digitized version of the classic *Dictionary of Icelandic* (Árnason, 2000), first published in 1963, as well as 20th century headwords from the AMI Written Language Archive (WLA, Ritmálssafn Orðabókar Háskólans[8]), a collection of citations created for a historical dictionary of Icelandic from the 16th century to modern times, which contains over 700,000 headwords. The third initial source was a book of personal names (Kvaran and Jónsson, 1991), containing about 4,800 personal names. The first version of the DMII contained 176,000 paradigms, mainly of vocabulary from these three sources. The additional material in later versions of the DMII comes from various other sources, and it is sometimes the result of cooperation on projects creating search engines capable of finding all inflectional forms of a word, by entering either the base form (the headword) or any inflectional form. This is true of the online version of the new translation of the *Bible* (2007), and the Icelandic telephone directory, a good source of named entities. These sources are, however, not to be relied upon for actual information on inflection, except for random forms, and it is only the published lexicographic work (*The Dictionary of Icelandic* and the book of names (Kvaran and Jónsson, 1991)) that contain a systematic coding for inflection, and only a partial one at that. It was therefore clear from the beginning that grammatical descriptions would be relied on, but the fact that these would prove to be incomprehensive was not immediately obvious.

The tradition in Icelandic dictionaries is to give certain inflectional forms as indicators of inflectional class, such as the genitive singular and nominative plural for nouns, either by showing the endings, e.g., *bátur*, (masc.) -s, -ar 'boat', or, in the case of wowel change, by showing the whole inflectional form, e.g., *köttur*, (masc.) *kattar, kettir* 'cat'. The remaining inflectional forms of nouns are very rarely shown in dictionaries, although some of these can be unpredictable, as in the masculine noun *bátur* 'boat', dative singular indefinite *báti* or *bát*, and dative singular definite *bátnum* (not **bátinum*) (cf. *köttur* 'cat', dat.sg.indef. *ketti*, dat.sg.def. *kettinum*). Information on the inflection in other word classes is also fragmentary, with the description of verbs usually confined to the principal parts, i.e., three or four inflectional forms, depending on inflectional class. The inflection of adjectives is very often omitted altogether, although it is not wholly predictable from the base forms. The dictionaries cover a large vocabulary, but they only give information on a part of the inflectional forms needed for complete paradigms.

The grammatical descriptions, on the other hand, show full paradigms of selected examples to give a survey of the system, i.e., they present the general structure 'top-down'.[9] This is true throughout the history of the description of Icelandic inflection, from the first one, which is fragmentary, usually referred to as *Grammaticæ islandicæ rudimenta*, first published in Copenhagen in 1651 (Jónsson, 1688), to the first definitive one, Rasmus Rask's *Vejledningen* (1811), up until now (cf. Kvaran, 2005). All the inflectional descriptions share the same characteristics, i.e., they present a set of generalized inflectional classes, mentioning exceptions at times.[10]

The grammar books are therefore not a good source on individual words, apart from the few selected examples, which usually are from the classic Icelandic core vocabulary. The emphasis on the core vocabulary means that data on loanwords, informal language, and slang is mostly absent from the sources, even though that is where changes to the system will first appear. Such material is ignored, perpaps for reasons of language purism,[11] even when such words seem to be fully adapted to the language, appearing in any syntactic context and being fully inflected, sometimes exhibiting major systematic differences from the traditional inflectional classes.

A case in point is the ending -i which the grammatical literature claims to be universal in the dative singular of neuter nouns, with the exceptions of four words. However, the data used for the DMII shows that the dative ending fluctuates between -i and -0 in multisyllabic neuter loanwords. This is the case for the the word *fennel* which is adopted from English, instead of the Icelandic version *fennika* (fem.), preferred by the purists. Even though some loanwords exhibiting this kind of fluctuation were adopted in the 18th century, they are still absent from the grammatical surveys. The word *arsenik* 'arsenic' is a case in point:[12]

[8] `http://arnastofnun.is/page/arnastofnun\_gagnasafn\_ritmal`

[9] The notable exception is Svavarsdóttir (1993), a monograph on the productivity of the inflectional classes of nouns, based on an empirical study of a corpus of 2.5 million running words.

[10] The most comprehensive one, *Islandsk grammatik* (Guðmundsson, 1922), proved to be immensely helpful, especially in the conjugation of verbs, but nevertheless it shares the characteristics of being a survey with the rest of the grammatical literature.

[11] There is a strong tradition of language purism in Iceland, i.e., a strong bias in favour of neologisms coined from Icelandic words in preference to loanwords.

[12] Citations from `http://timarit.is`, The National and

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

The lack of information in the two types of sources is therefore as follows: The dictionaries give partial information on quite a large vocabulary, but the grammatical descriptions give exhaustive information on a part of the vocabulary. Additional data is clearly needed.

5. Research for the DMII

It is of course only necessary to research possible ambiguous inflectional forms, but considerable research was (and is) needed to fill the above-mentioned gaps in the sources used for the DMII, using all the available sources at the AMI Department of Lexicography, i.e., the archives, citations in printed dictionaries, and digitized text collections, both at the AMI and at the National Library, as well as native speaker intuition, both from linguists and others. The users of the online DMII are also very generous with their opinions and suggestions for additions and improvement. At a pinch, Google is also used for reference, time-consuming though that may be. It still remains a fact that some problematic inflectional forms simply cannot be found anywhere, by any means.

To name an example, the word *Yggdrasill* (from Old Norse mythology, 'the great tree whose branches and roots extend through the universe') does not appear in the dative in any of the Old Icelandic sources. There are two possible dative forms, *Yggdrasil* and *Yggdrasli* and neither of them are attested in the literature. The second variant would be the regular inflection, but confusion with the neuter noun *drasl* (dat. *drasli*) 'rubbish' makes modern speakers cringe (or laugh), although the first variant is not quite acceptable either. This seems to make speakers avoid referring to a shop in today's Reykjavík named *Yggdrasill* in the dative, making do with syntactic context where another case can be used.[13] In the case of unattested inflectional forms, the choice is between a blank and an educated guess; both occur in the DMII. In the case of *Yggdrasill*, the first dative variant is shown (*Yggdrasil*), with a note in the online version stating that the form is unattested in the trustworthy sources.

Not all attested forms find their way into the DMII, although the purpose is description rather than prescription. Attested but totally unacceptable inflectional variants are not included in the DMII, such as the plural *fótar* of *fótur* 'foot', instead of *fætur*. (The plural *fótar* is sometimes heard in the speech of children and foreign learners, cf. English *foot*, pl. *feet*, not *foots*.) Such forms can often be excluded on the grounds of frequency, but the balance between description and prescription is a difficult one and the choices made can be subjective. For LT analysis, it might be better to include unacceptable variants, but the users of the online version would be up in arms to see them, as inclusion in the DMII is taken to be a kind of recognition of correctness.[14]

The limitations on the kind of research described here are of course the fact that the researcher has to rely on intuition to a great extent when deciding what to search for. It is simply not possible to use this method to search exhaustively in unannotated material, as the word forms themselves are ambiguous, both within paradigms and between lemmas. The identical nominative and accusative singular forms of the noun *akur* in Table 1. are examples of ambiguity within a paradigm, and the word form *minni* is an example of ambiguity between lemmas, as it appears in four different paradigms as 36 different inflectional forms, i.e., as the comparative of *lítill* 'small' (20 different tags), in the verb *minna* 'remind; remember' (10 different tags), in the neuter noun *minni* 'memory' (5 different tags), and as the feminine dative singular of the 1st person possessive pronoun *minn* 'my'.

The ambiguity is extensive. There are 2.8 million word forms (unique character strings) contained in the 5.8 million inflectional forms in the DMII, 1.8 million of the word forms are unique inflectional forms and thus unambiguous, but 1 million word forms are ambiguous. The figures for the ambiguity of inflectional forms within and between lemmas is shown in Table 4.[15]

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%

Table 4: Ambiguity of inflectional forms in the DMII.

An annotated corpus, such as the MÍM Corpus, is therefore an extremely important resource in the work on the DMII, both to resolve ambiguity and as a source of additional vocabulary.

6. The MÍM Corpus and the DMII

The MÍM Corpus is due to be completed later this year, and it will contain about 25 million running words. The first stage in comparing the vocabulary in the MÍM Corpus and the DMII is now in progress, with inclusion in the DMII in mind. Word forms from the ca. 17.7 million running words available from the MÍM Corpus when the process began have been compared to the DMII.[16] When all strings

University Library of Iceland's digital library of journals and newspaper texts.

[13] Examples of the word form *Yggdrasli* can be found on the web, often in disputes about the dative form.

[14] The solution in the DMII is data driven, which is much too liberal for some, but not descriptive enough for others. There are, however, extensive notes on the variants on the website to help the users cope. The DMII could therefore perhaps be said to be prescriptive, as it is used for guidance, but that prescription is by data, rather than by fiat. Still, the question of acceptability is an acute one.

[15] Figures from October 2011.

[16] It should be noted that the version of the MÍM Corpus used in the comparison contains additional material which will not be a

containing numerals, symbols and punctuation have been removed, the total number of tokens used in the comparison stands at 16,245,429.

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

Table 5: Tokens and unique tagged word forms in 1st batch of MÍM.

The lemmatization of the tagged word forms not found in the DMII has been checked and corrected manually. The corrections range from changes in the form of the lemma to changes of word class and/or gender of nouns. An example of both occurs in the lemmatization of the inflectional form *Miklagarði* with the lemma form *Miklagarð*, tagged as a feminine noun (fem.sg.acc.proper name). The correct lemma is *Mikligarður*, i.e., a masculine noun, with the nominative ending -ur, and a sound change in the stem.[17] It should be noted that the lemmatizer is used on PoS tagged text, and thus it inherits the mistakes made in the assignment of word class or the gender of nouns made by the tagger. This in turn is a source of mistakes in the assignment of the form of the lemma.

Of the 312,618 tagged word forms inspected at this stage of the work, 34.5% were found to be correctly tagged, both for the form of the lemma and word class. The number of inflectional forms in the major word classes (and gender for nouns) is shown in Table 6, where the first column of figures is the number of inflectional forms as tagged in the MÍM Corpus, the second one is the number of correctly assigned inflectional forms in the MÍM Corpus, and the last one is the result of the correction, i.e., the actual number of inflectional forms in the word class.

	MÍM	Correct	Result
n.neut.	70,770	31,942	54,317
n.masc.	142,630	42,910	64,709
n.fem.	78,908	46,207	64,499
adj.	23,835	11,713	18,345
verb	7,639	1,035	3,210

Table 6: Number of inflectional forms per word class, before and after correction.

60% of the word forms from MÍM not found in the DMII were found to be true Icelandic inflectional forms. These are the candidates for inclusion in the DMII. The remaining 40% were classified into a few categories, simultaneously with the correction of the lemmatization. The bulk of these uninflectable word forms and extraneous material is foreign words, 24.6% of the total, but other categories include errors (5.8%), abbreviations (1.6%), and computer-oriented strings (email addresses, urls, etc.), (0.7%).

This material has no direct relevance to the DMII, but it may come in useful in projects using the DMII data, such as context sensitive spelling correction. Hopefully, this data can also be of use in further work on the MÍM Corpus.

The preliminary results indicate that just over 125,000 paradigms should be added to the DMII. With an average number of inflectional forms per paradigm just exceeding 20, the necessary additions to the DMII could consist of over 2.5 million inflectional forms, bringing the total number of inflectional forms in the DMII to over 8.3 million. These figures are, however, preliminary only; the corrected list of lemmas from the MÍM Corpus has to be checked against the lemma list from the DMII again. This can not be said to be completed until the new material has been added to the DMII, as that process will serve to verify the lemmatization.

Work on the inclusion of the additional paradigms in the DMII is in progress. Inflectional class is assigned to each lemma, by comparison to previous DMII material, with the aid of a recently developed compound splitter.[18] Values are then assigned to the flags for grammatical categories for each lemma (cf. Table 1). The actual inflectional forms from the MÍM Corpus will be used for both processes, along with additional data from other sources at need. A revision of paradigms presently in the DMII will also take place, when the MÍM Corpus yields additional forms.

7. Data scarcity in a rich morpholgy

The part of the MÍM Corpus used in the project described here yields 737,856 word forms, and the estimated number of inflectional forms is just under 623,000 (84%). This is only a small part 5.8 million inflectional forms presently found in the DMII.[19] These figures give an indication of how large a corpus would have to be in order to be a sufficient base for a description (or a rule-system) of Icelandic inflection. A description of inflection based solely on the MÍM Corpus would be very meager indeed.[20] This problem is no surprise to Icelandic lexicographers, who have always had to cope with a similar problem in a different context (Pind et al., 1993), as the same kind of scarcity problem is seen in the search for inflectional forms as in the search for words in specific syntactic context.

The proposed addition of a further 125,000 paradigms to the DMII, on the basis of the comparison with the MÍM Corpus, brings the total of inflectional forms in the DMII to 8.3 million, i.e., the 623,000 forms from the Corpus spawn

part of the final version. The tagging is not the final product either, with some texts not tagged with the final combination of taggers described in Loftsson et al. (2010).

[17] *Mikligarður* is the Old Norse name of Constantinople, sometimes still used of Istanbul in Modern Icelandic. The word is a compound, from *mikill* 'great' and *garður* 'seat (of a king); city'.

[18] Work in progress, by Jón Friðrik Daðason, in cooperation with the AMI.

[19] The total number of lemmas in the MÍM Corpus cannot be compared to the number of lemmas in the DMII, as the comparison of word forms described above was defined to word forms not present in the DMII. The lemmatization of the remaining word forms, i.e., those appearing both in the MÍM Corpus and the DMII, is known to contain too many errors to be meaningful at this stage.

[20] For comparison, the *Icelandic Frequency Dictionary* (Pind et al., 1991), which is based on a corpus of 500,000 running words, contains 31,876 lemmas and 59,343 inflectional forms.

a estimated 2.5 million new forms. This would bring the total number of paradigms in the DMII to 395,000, still far short of the over 700,000 headwords found in the largest lexicographic archive (WLA) at the AMI. The indication is, therefore, that the lexicographic data and the corpus complement each other as valuable sources.

8. Conclusion

As pointed out in the introduction, the DMII was intitially made mainly for two purposes, i.e., to produce data for LT use and as reference material for the general public. In the process of the work, the role of the DMII in language research has become increasingly important. In spite of the reputation of Icelandic as a relatively well-researched language, with centuries of history of morphological description, it turned out that there was and still is a lot of work to be done in the field. The fact that the inflectional system is both complex and irregular, with enormous fluctuation between variant forms and inflectional classes, made it necessary to produce a complete set of paradigms, which also allows for notes to be included on individual words at need. The notes contain data on the underlying research, and indications on usage. The usage notes are published on the website, and they are aimed at the general public, as there can often be semantic or stylistic restrictions on the choice of variants.

The production of a rule system of Icelandic inflection is gradually becoming more feasible, as the scope of the DMII is expanded. A morphological analyzer based on the DMII would also be useful, as the DMII will of course never be all-inclusive. It should however be emphasized that the DMII is still work in progress.

The data from the DMII is available online for LT projects on the AMI website, free of charge. Conditions on the use of the data are published on the website.

9. Acknowledgements

The project was initially financed by The Language Technology Program of the Ministry of Education, Science and Culture which ran during the years 2002–2004. The creation of the DMII database was financed by Technology Development Fund in 2005–2006. The company Já has cooperated with the AMI since 2008 on maintaining the database, as well as giving a grant in order to facilitate the use of the DMII for LT purposes. The author would like to thank numerous contributors to the project since 2002, including the well-wishers contributing additions and corrections to the online version by email. Special thanks are due to Þórdís Úlfarsdóttir, who worked on the project for two years at a critical time, and to Sigrún Helgadóttir, from the MÍM Corpus, and Jón Friðrik Daðason, M.Sc. student in Computer Science at the University of Iceland, without whose help the comparison of data from the DMII and MÍM would not have been made. Thanks are also due to the anonymous SALTMIL reviewers for their helpful comments.

10. References

M. Árnason. *Íslensk orðabók [Dictionary of Icelandic]*. 2000. Edda hf., Reykjavík, 3rd edition, electronic version.

Biblían [The Bible]. 2007. Hið íslenska biblíufélag, Reykjavík, 11th edition.

V. Guðmundsson. *Islandsk grammatik [Grammar of Icelandic]*. 1922. Hagerup, Copenhagen.

S. Helgadóttir, Á. Svavarsdóttir, E. Rögnvaldsson, K. Bjarnadóttir, and H. Loftsson. 2012. The Tagged Icelandic Corpus (MÍM). In *Proceedings of "Language Technology for Normalization of Less-Resourced Languages", workshop at the 8th International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey. Submitted.

R. Jónsson. *Grammaticæ islandicæ rudimenta*. 1688. E theatro Sheldoniano, Oxford, 2nd edition [1st edition, 1651].

G. Kvaran. *Íslensk tunga 2. Orð. Handbók um beygingar- og orðmyndunarfræði*. 2005. Almenna bókafélagið, Reykjavík.

G. Kvaran and S. Jónsson. *Nöfn Íslendinga [The Names of the Icelanders]*. 1991. Heimskringla, Reykjavík, 1st edition.

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

J. Pind, F. Magnússon, and S. Briem. 1991. *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavík.

J. Pind, K. Bjarnadóttir, J. H. Jónsson, G. Kvaran, F. Magnússon, and Á. Svavarsdóttir. 1993. Using a Computer Corpus to Supplement a Citation Collection for a Historical Dictionary. *International Journal of Lexicography*, 6(1):1–18.

R. K. Rask. 1811. *Vejledningen til det Islandske eller gamle nordiske Sprog*. Schubothes Forlag, Copenhagen.

E. Rögnvaldsson, H. Loftsson, K. Bjarnadóttir, S. Helgadóttir, A. B. Nikulásdóttir, M. Whelpton, and A. K. Ingason. 2009. Icelandic Language Resources and Technology: Status and Prospects. In R. Domeij, K. Koskenniemi, S. Krauwer, B. Maegaard, E. Rögnvaldsson, and K. de Smedt, editors, *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*. Odense, Denmark.

Á. Svavarsdóttir. *Beygingakerfi nafnorða í nútímaíslensku [The Inflectional System of Nouns in Modern Icelandic]*. 1993. Málvísindastofnun Háskóla Íslands, Reykjavík.