MIM-GOLD

Starkaður Barkarson, Þórdís Dröfn Andrésdóttir, Hildur Hafsteinsdóttir, Árni Davíð Magnússon, Kristján Rúnarsson, Steinþór Steingrímsson, Haukur Páll Jónsson, Hrafn Loftsson, Einar Freyr Sigurðsson, Eiríkur Rögnvaldsson, Sigrún Helgadóttir
June 1, 2021

1 Introduction

This paper describes the work on MIM-GOLD step by step. The primary changes of the tagset for grammatical tagging of Icelandic texts throughout the years are explained. A detailed description of the latest changes made on the tagset is available in Appendix I.

2 Changes to the tagset and work on MIM-GOLD

The revised tagset (see chapter 2.6 for further details) builds upon a tagging scheme created for the Icelandic Frequency Dictionary (IFD) in 1991 (Pindet al., 1991). The tagging scheme has been changed as described below. Even more changes were made during the revision for version 20.05. For instance, abbreviations and shortened forms were put in a separate category. A new category was created for symbols which have become more common in written language, especially on the internet. Changes were made to how foreign words are analyzed, i.a. More information on these changes is found in chapter 3.

The Tagged Icelandic Corpus (MÍM) was published in 2013. The corpus contains about 25 million running words of texts written during the first decade of the 21st century.

While MÍM was being compiled, about one million tokens were sampled from 13 of the 23 domains of MÍM. The new corpus should replace the corpus of the Icelandic Frequency Dictionary (IFD) as a gold standard for the training of data-driven taggers for Icelandic.

In 2013 version 0.9 of MIM-GOLD was published. In 2018, access to version 1.0 was granted. The development of MIM-GOLD is described below. The process is divided into 7 phases, numbered from 0 to 6.

2.1 Phase 0

Work on MIM-GOLD commenced in the summer of 2009 when a grant was secured from the Student Innovation Fund to hire a student to start the project. The texts were sampled at the Árni Magnússon Institute for Icelandic Studies and the student under the supervision of Hrafn Loftsson at Reykjavík University developed a system for tagging the texts. The texts were tokenized with a tokenizer that is a part of the IceNLP system. The texts were then tagged with five taggers: fnTBL, MXPOST, IceTagger, Bidir and TnT (Loftsson et al., 2010). The tool CombiTagger was then used to vote between the proposed tags. A method was used that chooses the tag that most taggers suggested. The taggers were trained on the corpus of the Icelandic Frequency Dictionary (IFD). The tagset of the IFD was therefore used.

During the winter of 2009–2010 a search for systematic errors in the MIM-GOLD corpus was performed. Noun phrase (NP), prepositional phrase (PP) and verb phrase (VP) error detection programs described by Loftsson (2009) were used. A large proportion of the errors detected were checked manually and errors corrected. Tagging accuracy was then estimated by inspecting every 100th word. A tag is correct if the whole tagstring (consisting of up to 6 characters) is correct. Mean tagging accuracy was estimated as 92.3%, ranging between 87.6 and 95.5% depending on text domain (Loftsson et al., 2010). This part of the project also received a contribution from a grant from the Icelandic Research Fund3.

2.2 Phase 1

During the summer of 2010, another grant was secured from the Student Innovation fund4 to employ a student to manually check and correct tags of all the words in MIM-GOLD. The first job was to finish checking errors found during Phase 0 that had not been corrected (texts from Morgunblaðið). Work on checking texts from printed books was also started. The student was then hired part time during term time and during 2010–2011 all the words in MIM-GOLD were manually checked and corrected. Version 0.9 of MIM-GOLD that was made available on this website in 2013 contains the files after this correction phase. Mean accuracy was estimated as before by inspecting the tag for every 100th word. Mean accuracy was estimated as 96.4%, ranging between 89.9% and 98.5% depending on text domain (Helgadóttir et al., 2014). The project also got a contribution from META-NORD and the Ministry of Education, Science and Culture.

2.3 Phase 2

The next correction phase started at the end of 2012. The corpus was first tagged automatically with the tagger IceTagger which is a part of the IceNLP software. A script was written that compares the tags output by IceTagger with the (presumed) correct tags in the corpus. If a difference was found the discrepancy was marked as an error candidate. A second student was employed during the summer of 2013 and part-time after that to inspect manually the error candidates. For each error candidate, the student was instructed to i) select the tag in the corpus; or ii) select the tag proposed by IceTagger; or iii) select a new correct tag when neither IceTagger nor the corpus contained the correct tag. After about 80% of the texts had been checked and corrected tagging accuracy was estimated as 99.6%, ranging between 99.5 and 100.0% depending on text domain (Helgadóttir et al., 2014). One more student was employed in late 2013 to finish checking and correcting the tags. That work was finished in 2014. Tagging accuracy was not estimated at the end of this phase. This part of the project was supported in part by META-NORD and the Ministry of Education, Science and Culture.

2.4 Phase 3

Steinþór Steingrímsson, Sigrún Helgadóttir and Eiríkur Rögnvaldsson experimented in 2015 with training the tagger Stagger (Östling, 2012) on the IFD and MIM-GOLD (Steingrímsson et al., 2015). Hrafn Loftsson and Robert Östling experimented in 2013 with developing a tagger for Icelandic by training and testing Stagger on the IFD and obtained 93,84% accuracy (Loftsson and Östling, 2013). Since this was the best result obtained so far with tagging Icelandic text it was decided to test Stagger on MIM-GOLD. By comparing the accuracy obtained when training and testing Stagger on the IFD and on MIM-GOLD it was clear that there were still a number of inconsistencies and incorrect tags in MIM-GOLD (Steingrímsson et al., 2015). For the experiment a version of MIM-GOLD after the completion of Phase 2 was used. The experiment with training and testing Stagger on IFD reported by Loftsson and Östling (2013) was repeated for MIM-GOLD by using linguistic features (LF) and the unknown word guesser IceMorphy (part of the IceNLP software). An extended lexicon based on the Database of Icelandic Inflection (BÍN) was added. By applying ten-fold cross-validation 92.76% accuracy was obtained for MIM-GOLD. As a result of this outcome it was decided to work further on reducing the number of errors and inconsistencies in MIM-GOLD. Lists of inconsistencies and errorswere made and students were employed to check them manually. The tagset was also modified slightly. Work on this phase was completed in 2017. This part of the project was funded by the Institute of Linguistics at the University of Iceland and the Icelandic Ministry of Education, Science and Culture.

2.5 Phase 4

Starkaður Barkarson obtained the data of the MIM-GOLD after Phase 3 was completed and trained Stagger on the texts (Barkarson, 2017). Tagging accuracy was not estimated after Phase 3 by inspecting a sample as had been done after previous correction phases. Barkarson repeated the experiment performed by Steinþór Steigrímsson, Sigrún Helgadóttir and Eiríkur Rögnvaldsson in 2015. He performed a comparable ten-fold cross-validation on MIM-GOLD and obtained 92.74% accuracy.

Despite the corrections made to MIM-GOLD tagging accuracy did not seem to increase. To make sure that the experiments were completely comparable the experiment performed by Steingrímsson et al. (2015) was repeated as far as possible. Same version of MIM-GOLD (before Phase 3) was used and same division into training and testing sets. Data for Database of Icelandic Inflection (BÍN) were not completely comparable since now a later version was used. Barkarson obtained 92.41% accuracy by using BIN and IceMorphy as compared to 92.76% in the experiment performed by Steinþór Steingrímsson and colleagues. Barkarson therefore claims that corrections made to MIM-GOLD resulted in an increase in accuracy of 0.30 percentage points. He believes that the reason for the difference may be found in the set of words and word endings that was available to IceMorphy since there is a large difference in accuracy of unknown words (just under 15%) but a small difference in accuracy of known words (0.09%) (Barkarson, 2017).

2.6 Phase 5

The morphosyntactic tagset for Icelandic was revised at the end of 2019 and a new version published under the name Tagset MIM-GOLD 2.0. This phase was a part of the Icelandic Language Technology Program 2019-2023, which was funded by the Ministry of Education, Science and Culture. A detailed description of the changes that were made on the tagset in this phase can be found in Appendix I. Next, the revision of the Icelandic gold standards, IFD and MIM-GOLD, commenced, taking the revised tagset into consideration. The tags were automatically converted to the new tagset and the primary changes were reviewed and corrected as needed so the tagging could be as accurate as possible. Þórdís Dröfn Andrésdóttir and Hildur Hafsteinsdóttir worked on reviewing and correcting the tags. Decisions in cases of doubt were made in consultation with Starkaður Barkarson, Einar Freyr Sigurðsson, Steinþór Steingrímsson and Eiríkur Rögnvaldsson. In the spring of 2020, the tagset was published alongside the revised edition of the Icelandic gold standard, MIM-GOLD 20.05. We also describe the revision of IFD, which is published separately. MIM-IFD is a combined package containing both gold standards and is published with IFD 20.05. The version numbers are in accordance with version number standards for published content under the language technology program. The gold standards are published simultaneously, in ten sections, to use for training and testing on grammar taggers.

2.7 Phase 6

In the beginning of 2021, the addition of lemmas to MIM-GOLD commenced and a new version was published in June the same year. The text had been lemmatized with *Nefnir* (Svanhvít Lilja Ingólfsdóttir o.fl., 2019) a few years earlier and some manual correction had taken place. Árni Davíð Magnússon and Kristján Rúnarsson worked on reviewing and correcting the lemmas. During the process, a few tags were corrected. This phase was a part of the Icelandic Language Technology Program 2019-2023, which was funded by the Ministry of Education, Science and Culture

2.8 Modified tagset

To simplify grammatical analysis and reduce inconsistencies in tagging, the tagset of the IFD was slightly modified during correction phases of MIM-GOLD. These changes were made:

- Foreign names were originally tagged as proper nouns. During Phase 3, they were tagged as foreign words (**e**). Steingrímsson et al. (2015).
- In the IFD, function words preceding *að* were classified as adverbs (**aa**). During Phase 2 on the other hand, they were classified as prepositions when followed by a complement clause. Thus, the word *til* in the sentence "Hann hljóp til að komast fyrr heim" is classified as a preposition governing genitive case (**ae**) (Helgadóttir et al., 2014; Steingrímsson et al., 2015; Barkarson, 2017).
- Further classification of proper nouns was discontinued during Phase 3. Tags of all proper nouns now end in **-s**, instead of **-m** (person names), **-ö** (location names) and **-s** (other proper nouns). Number of tags is reduced by 68 (Steingrímsson et al., 2015)
- During Phase 3, **v** was adopted as a tag for e-mail addresses and web addresses (Steingrímsson et al., 2015)
- During Phase 3, as was adopted as a tag for abbreviations. In the IFD tagset, abbreviations were broken up into individual words and each letter tagged as the word it stood for (Steingrímsson et al., 2015).
- During Phase 3, it was decided that all number constants that were tagged as cardinals (**tf...**) should be given the tag **ta** and not analyzed further according to gender, number and case, as is done when numbers are written with alphabetic characters (Steingrímsson et al., 2015).
- During Phase 5, a new tagset was introduced, as described in chapter 2.6. A detailed description of the changes made can be found in Appendix I.

3 Instructions on gold standard tagging

This chapter describes the decisions made on tagging the gold standards. To ensure consistency in tagging, all decisions were documented. Following these decisions could help potential future tagging of texts to be in accordance with the present published corpus.

3.1 Abbreviations and shortened forms

5.1.1 Abbreviation or shortened form?

- If a word form stands only for one word and is the former part of the word (*lögg.* = *löggiltur* 'certified', *hæstv.* = *hæstvirtur* 'honorable'), or if a word is a compound word, put together of two word forms (*lög.stj.* = *lögreglustjóri* 'police commissioner', *framkv.stj.* = *framkvæmdastjóri* 'manager'), and the word form consists of three or more letters, the word form is a shortened form and not an abbreviation.
- Therefore, *lögg.*, *hæstv.*, *lög.stj.* and *framkv.stj.* are shortened forms.

5.1.2 Foreign abbreviations that are not proper nouns

- Foreign shortenings and abbreviations (that do not stand for proper nouns) are **e**.

3.2 Icelandic proper nouns

- In proper nouns consisting of more than one word, such as *Bóksala stúdenta* 'the Students' bookstore', only the first word is tagged as a proper noun.

  Bóksala **nven-s**
  stúdenta **nkfe**

3.3 Foreign proper nouns

5.3.1 General guidelines

- **Person names:** are always tagged as **n----s**. If a person has more than one name, each name is tagged with **n----s**. (e.g. all parts of *Alessandro Del Piero* are tagged as **n----s**).
- **Location names:** are always tagged as **n----s**.
- **One-word names of institutions and companies:** are tagged as **n----s**.
- **Long, compound proper nouns:** It is generally the rule that only the first word is tagged as **n----s** and the rest as foreign words, **e**, except those who stand as proper nouns on their own. Names of persons and locations do not fall under this rule, see above.

5.3.2 Special cases

- Car brand subclasses (e.g. *Skoda **Superb***, *VW **Passat***, *Renault **Megane***) are tagged as **n----s**, as they are used independently (separately from the brand name) and are proper nouns on their own.
  - If another word follows the subclass (e.g. *Renault Megane **Saloon***), it is tagged as **e**.
- Names of foreign sports clubs named after cities or other places (e.g. *Los Angeles **Lakers***, *New York **Knicks***, *Utah **Jazz***) are tagged as **n----s**, as they are often used independently and they are therefore proper names on their own.
- Foreign titles (e.g. ***Dame** Judi Dench*, ***Mr** Feather*, ***Major** Miriam Oskarsdóttir*) are tagged as **e** as they are not a part of the proper noun which follows. However, titles of the same sort are tagged as **n----s** when they are the first part of a proper noun, such as in movie titles (e.g. ***Mr** Deeds*, ***Mrs** Doubtfire*).
- Foreign, compound-word names of companies, bands, book titles, movie titles, art titles, conferences and fairs are tagged as follows:

Y **n----s**
tu **e**
mamá **e**
también **e**

- If a proper name is contained within a proper name like this one, it gets the tag **n----s**.

3.4 Icelandic or foreign?

- Foreign words which have been adapted to the Icelandic inflectional system (such as *Steinwayinum* and *Bösendorferinn*) are tagged as Icelandic words:

  Steinwayinum **nkeþgs**
  Bösendorferinn **nkengs**

3.5 Other

- Symbols: **m** or **pa**
  - If a traditional punctuation mark stands for a word it is **m**.
    Example: '-' in *2-3 hours*, '/' in *km/hour*.

References

Barkarson, Starkaður. 2017. Þjálfun málfræðimarkarans *Stagger* með nýjum gullstaðli. Master's thesis, The University of Iceland, Reykjavík. http://hdl.handle.net/1946/29474.

Helgadóttir, Sigrún, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2014. Correcting Errors in a New Gold Standard for Tagging Icelandic Text. In *Proceedings of the* 9*th* *International Conference on Language Resources and Evaluation*, LREC 2014. Reykjavik, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/summaries/677.html.

Loftsson, Hrafn. 2009. Correcting a POS-tagged corpus using three complementary methods. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 523–531. Athens, Greece: Association for Computational Linguistics. https://www.aclweb.org/anthology/E09-1060.

Loftsson, Hrafn, and Robert Östling. 2013. Tagging a morphologically complex language using an averaged perceptron tagger: The case of Icelandic. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, 105–119. Oslo, Norway: Linköping University Electronic Press, Sweden. https://www.aclweb.org/anthology/W13-5613.

Loftsson, Hrafn, Jökull H. Yngvason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. In *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, LREC 2010. Valetta, Malta. https://www.isca-speech.org/archive_open/saltmil/SALTMIL2010_Proceedings.pdf#page=57.

Pind, Jörgen, Friðrik Magnússon, and Stefán Briem, eds. 1991. *Íslensk orðtíðnibók*. Reykjavík: Orðabók Háskólans.

Steingrímsson, Steinþór, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2015. Analysing Inconsistencies and Errors in PoS Tagging in two Icelandic Gold Standards. In *Proceedings of the* 20*th* *Nordic Conference of Computational Linguistics*, NODALIDA 2015, 287–291. Vilnius, Lithuania. https://www.aclweb.org/anthology/W15-1838.

Steingrímsson, Steinþór, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP 2019. Varna, Bulgaria.

Svanhvít Lilja Ingólfsdóttir, Hrafn Loftsson, Jón Friðrik Daðason, Kristín Bjarnadóttir. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. *Proceedings of the 22*nd *Nordic Conference on Computational Linguistics*, 310–315. Turku, Finland.

Östling, Robert. 2012. Stagger: A modern POS tagger for Swedish. In *Proceedings of the Swedish Language Technology Conference, SLTC*. Lund, Sweden.

Appendix I

1 Tagset changes in Phase 5

More changes were made to the tagset in this version as mentioned in the introduction. They are explained here.

1.1 Prepositions

- **ao**, **aþ** and **ae** become **af** (adverb that assigns case).

1.2 Verbs

- What was tagged as the supine (**ssg** or **ssm**) is now tagged as a past participle: verb(part.)-past.-active-neut.-sing.-nom. (**sþghen**) or verb(part.)-past.-middle-neut.-sing.-nom. (**sþmhen**).

1.3 Nouns

- A tag for unspecified gender (**nx...**) was removed from the tagset to be replaced with **-** (**n-**…).

1.4 Foreign words

- Foreign proper nouns will be tagged as **n----s**, i.e. nouns that are not tagged for gender, grammatical number or case and do not have a definitive article.
- Foreign abbreviations will be tagged in the same way if they are the equivalent of a proper noun (e.g., *CIA*, *NATO*, *LFC* and *KFC*).

1.5 Abbreviations and shortened forms

- Abbreviations and shortened forms are in a seperate category where the tagset starts with **k**.
- Text abbreviations are tagged as **ks**. If the abbreviation is a a name (e.g., *KR*, *VR* or *ÓRG*) it is tagged as **n----s**. This applies both to foreign and to Icelandic abbreviations (see above).
- Shortened forms are tagged as **kt**. Shortened forms are, e.g., *lögg.* for *löggiltur* ‘certified’, *hæstv.* for *hæstvirtur* ‘honorable’, i.e., where a latter part of the word is omitted in writing and replaced with a full stop.
- Furthermore, *fót-* in *fót- og handbolti* ‘foot- and handball’ is tagged as **ks**.

1.6 Punctuation

- **pl**, at the end of a sentance: .!? (always)
- **pk**, comma: , ; (except if , is used as a quotation mark)
- **pg**, quotation marks: « » „ “ ‟ ” ” , ‘ (except if , is a comma or ’ is an apostrophe)
- **pa**, other punctuation marks: ( ) { } _ : - – — … (plus all that did not fall into the aforementioned categories)

1.7 Multiple punctuation marks

- Two or more full stops (e.g., …), question marks (e.g., ??) or exclamation marks (e.g., !!!) in a row, or a combination of question marks and exclamation marks (e.g., !?!) is tagged together as **pa**.
- Other punctuation marks are each tagged seperately.

1.8 Symbols

- All symbols are tagged as **m**

  - math symbols: + − × ÷ = < > [ ]
  - emojis: :) ♥ …
  - other symbols: $ % § © •

- Symbols can be defined as most of which that does not include a letter, or a number, and is not a punctuation mark – except when two or more punctuation marks in a row form an emoji.
- A symbol can be replaced with a word (e.g., $ = dollar, + = plus).

Tagset MIM-GULL 2.0		
Column	**Category**	**Analytical symbol** – information
1	Word class	**n-noun**
2	Gender	**k**-masculine, **v**-feminine, **h**-neuter
3	Number	**e**-singular, **f**-plural
4	Case	**n**-nominative, **o**-accusative, **þ**-dative, **e**-genitive
5	Article	**g**-with suffixed definite article
6	Proper noun	**s**-proper noun
1	Word class	**l-adjective**
2	Gender	**k**-masculine, **v**-feminine, **h**-neuter
3	Number	**e**-singular, **f**-plural
4	Case	**n**-nominative, **o**-accusative, **þ**-dative**,** **e**-genitive
5	Declesion	**s**-strong declension, **v**-weak declension, **o**-indeclinable
6	Degree	**f**-positive**, m**-comparative**, e**-superlative
1	Word class	**f-pronoun**
2	Subcategory	**a**-demonstrative pronoun, **b**-indefinte demonstrative pronoun, **e**-possessive pronoun, **o**-indefinite pronoun, **p**-personal pronoun, **s**-interrogative pronoun, **t**-relative pronoun
3	Gender/Person	**k**-masculine, **v**-feminine, **h**-neuter**/1**-1. pers., **2**-2. pers.
4	Number	**e**-singular, **f**-plural
5	Case	**n**-nominative, **o**-accusative, **þ**-dative, **e**-genitive
1	Word class	**g-article**
2	Gender	**k**-masculine, **v**-feminine, **h**-neuter
3	Number	**e**-singular, **f**-plural
4	Case	**n**-nominative, **o**-accusative, **þ**-dative, **e**-genitive
1	Word class	**t-numeral**
2	Category	**f**-cardinal number, **a**-date and other indeclinable numbers, **p**-percentage, **o**-number which precedes other numerals
3	Gender	**k**-masculine, **v**-feminine, **h**-neuter
4	Number	**e**-singular, **f**-plural
5	Case	**n**-nominative, **o**-accusative, **þ**-dative, **e**-genitive
1	Word class	**s-verb** (except for past participle)
2	Mood	**n**-infinitive, **b**-imperative, **f**-indicative, **v**-subjunctive, **l**-present participle
3	Voice	**g**-active, **m**-middle
4	Person	**1**-1st person, **2**-2nd person, **3**-3rd person
5	Number	**e**-singular, **f**-plural
6	Tense	**n**-present, **þ**-past
1	Word class	**s-verb** (past participle)
2	Mood	**þ**-past participle
3	Voice	**g**-active, **m**-middle
4	Gender	**k**-masculine, **v**-feminine, **h**-neuter
5	Number	**e**-singular, **f**-plural
6	Case	**n**-nominative, **o**-accusative
1	Word class	**a-adverb**
2	Category/case governor	**a**-does not govern case, **f**-governs case, **u**-exclamation
3	Degree	**m**-comparative, **e**-superlative
1	Word class	**c-conjunction**
2	Category	**n**-sign of infinitive, **t**-relative conjunction
1	Word class	**k-abbreviation**
2	Category	**s**-abbreviation, **t**-short form
1	Word class	**e-foreign word**
1	Word class	**x-unanalysed word**
1	Word class	**v-e-mail, web address**
1	Word class	**p-punctuation mark**
2	Category	**l**-end of sentence, **k**-comma, **g**-quotes, **a**-others
1	Word class	**m-symbo**l