Developing a Faroese PoS-tagging solution using Icelandic methods

Hinrik Hafsteinsson and Anton Karl Ingason
University of Iceland
Reykjavík, Iceland
{hih43,antoni}@hi.is

Abstract

We describe the development of a dedicated, high-accuracy part-of-speech (PoS) tagging solution for Faroese, a North Germanic language with about 50,000 speakers. To achieve this, a state-of-the-art neural PoS tagger for Icelandic, *ABLTagger*, was trained on a 100,000 word PoS-tagged corpus for Faroese, standardised with methods previously applied to Icelandic corpora. This tagger was supplemented with a novel Experimental Database of Faroese Inflection (EDFM), which contains morphological information on 67,488 Faroese words with about one million inflectional forms. This approach produced a PoS-tagging model for Faroese which achieves a 91.40% overall accuracy when evaluated with 10-fold cross validation, which is currently the highest reported accuracy for a dedicated Faroese PoS-tagger. The tagging model, morphological database, proposed revised PoS tagset for Faroese as well as a revised and standardised PoS tagged corpus are all presented as products of this project and are made available for use in further research in Faroese language technology.

1 Introduction

This paper describes the developement of a high-accuracy Part-of-Speech (PoS) tagging solution for Faroese, a North Germanic spoken by about 50,000 people in the Faroe Islands, an autonomous territory in the Kingdom of Denmark. Limited research has been performed on such an implementation for the language and as PoS taggers are fundamental in various further implementations in natural language processing (NLP) and linguistic research, the need for new research is apparent. In the current project, as a basis for a Faroese PoS tagger, a state-of-the-art bi-LSTM neural PoS-tagging system for Icelandic, *ABLTagger* (Steingrímsson et al., 2019), is used as a foundation to build on, in addition to various methods used in Icelandic NLP research. The reason why these Icelandic resources may also be applied to Faroese is the extensive grammatical similarities between the two languages. These similarities are especially apparent in morphology, as both languages retain grammatical categories and nuances not apparent in related languages.

In contrast with Faroese, the last two decades have seen broad gains for Icelandic in the field of language technology (LT), producing tools and databases which have enabled both various new technical implementations for the language and new opportunities in linguistic research. With the grammatical similarities of the two languages in mind, similar gains should be possible for Faroese.

Using Icelandic NLP tools and methods, a Faroese PoS-tagging model was produced which achieves an overall tagging accuracy of 91.40%. This is considerably higher than the previous dedicated PoS tagger for Faroese which achieved 87.00% using a similar tagset and trained on the same corpus, the Faroese Sosialurin corpus, which contains news articles, totalling about 100,000 words (Hansen et al., 2004). Furthermore, the current project produced various data set innovations which could prove useful in further NLP research for Faroese. These include a proposed revised tagging scheme for Faroese which is optimised for high-accuracy PoS-tagging and a standardised version of hand-corrected PoS tagged corpus for use in NLP projects. In addition to this, the project produced a novel inflection database for Faroese, the Experimental Database of Faroese Morphology (EDFM), which contains detailed morphological information on 67,488 Faroese words, modelled on the Dictionary of Icelandic Morphology (Bjarnadóttir, 2012; Bjarnadóttir et al., 2019). These data sets have been made available online[1] for fur-

[1] See: https://github.com/hinrikur/far-ABLTagger.

Proceedings of the 17th International Conference on Natural Language Processing, pages 481–490
Patna, India, December 18 - 21, 2020.

ther development and research in Faroese language technology.

Section 2 outlines the ABLTagger PoS-tagging system and previous work on Faroese LT relevant to the current project, along with discussing the applicability and incentives to use Icelandic materials and methods to develop LT solutions for Faroese. Section 3 describes the collection and preparation of the materials used to implement ABLTagger for Faroese. Section 4 describes the training and evaluation of the Faroese ABLTagger model and Section 5 discusses the current applicability of the Faroese PoS tagging model and the next steps in improving it as a whole. Section 6 concludes.

2 Background

2.1 The ABLTagger system

The current project draws inspiration from the ABLTagger experiment for Icelandic (Steingrímsson et al., 2019), which produced a Bi-directional LSTM PoS tagger which uses a morphological database to achieve high accuracy tagging of the language, using a fine-grained tagset. When trained on a hand-corrected corpus, the *IFD* corpus (Pind et al., 1991), a total of about 500,000 words, the ABLTagger system achieved an overall tagging accuracy of 94.17%, making it the state-of-the-art PoS-tagging implementation for Icelandic.

The ABLTagger system for Icelandic uses a fine-grained tagset of about 600 PoS tags, originally introduced in the aforementioned IFD corpus and revised in further research, notably in the MIM-GOLD corpus (Barkarson et al., 2020). An excerpt of this tagging scheme is shown in Table 1. In the tagging scheme, each token receives one tag string. Each tag string contains a series of symbols, each containing important grammatical information on the token, e.g., case, number, tense and grammatical gender.

The morphological component of the ABLTagger system consists of a morphological database for the language being tagged. For Icelandic, Steingrímsson et al. used the Databse of Icelandic Morphology, *DIM* (Bjarnadóttir, 2012; Bjarnadóttir et al., 2019), in the so-called DIM basic format.[2] *DIM* contains around 290,000 inflectional paradigms with over 5.8 million inflectional forms, aiming to be a descriptive resource for Icelandic. This database is freely available under a CC BY-SA 4.0 license in standardised formats and is for the most part manually corrected.

Providing these two components of the ABLTagger system, the PoS-tagged corpus and morphological database, is essential for implementing the system for a new language and is the main focus of the current project.

2.2 Previous work on Faroese

Although extensive research in Faroese LT is scant, some research has taken place in the past decades. Experiments have been performed in machine-parsing Faroese text using transfer learning with Icelandic data (Ingason et al., 2014) and work on a finite-state based grammatical analyser for Faroese is ongoing (Trosterud, 2009).

Notably, the Sosialurin corpus project (Hansen et al., 2004) consisted of the formulation of a fine-grained PoS tagging scheme for Faroese and the compilation of a hand-corrected, PoS-tagged text corpus using the tagset. This corpus was used to train the TnT tagger (Brants, 2000), which achieved an overall tagging accuracy of 87.0% at the time. Furthermore, the Sosialurin corpus and further machine-tagged text have been made accessible for linguistic research on the CorpusEye (Bick et al., 2020) website.[3] These resources, the Sosialurin corpus and its corresponding tagging scheme, are instrumental for the current project, as they provide the training material for the ABLTagger system.

2.3 Applicability of Icelandic methods for Faroese

The foremost incentive for using Icelandic NLP tools for Faroese is the scarcity of NLP implementations for the latter. A handful of previous research exists, but very few extensive databases and ready-to-use software are available for the language at large. As a result, the amount of available digital language resources for Faroese is limited. In contrast, the last two decades have seen great gains for Icelandic in the field of language technology (Nikulásdóttir et al., 2017). This has produced tools and databases which enable both new technical implementations for the language and new opportunities in linguistic research, both of which would also be beneficial for Faroese.

The fundamental reason that makes Icelandic NLP implementations applicable for Faroese are the grammatical similarities between the two lan-

[2] See: https://bin.arnastofnun.is/DMII/LTdata/s-format/

[3] See: https://corp.hum.sdu.dk/.

Token	Lemma	PoS tag	Explanation
Ég	ég	fp1en	**f**: pronoun; **p**: personal; **1**: 1st person; **e**: singular; **n**: nominative;
stökk	stökkva	sfg1eþ	**s**: verb; **f**: indicative; **g**: active; **1**: 1st person; **e**: singular; **þ**: past tense
á	á	aa	**a**: adverb; **a**: doesn't govern case;
eftir	eftir	aþ	**a**: adverb; **þ**: governs dative;
strætó	strætó	nkeþ	**n**: noun; **k**: masculine; **e**: singular; **þ**: dative;
og	og	c	**c**: conjunction;
veifaði	veifa	sfg1eþ	**s**: verb; **f**: indicative; **g**: active; **1**: 1st person; **e**: singular; **þ**: past tense

Table 1: Excerpt from the *IFD* corpus, with explanations. The IFD corpus was used to train the ABLTagger system for Icelandic by Steingrímsson et al. (2019) and as a basis for the original Faroese tagging scheme by Hansen et al. (2004).

guages. These similarities are especially apparent in morphology, as both languages retain grammatical categories not apparent other similar languages, e.g., four grammatical cases for nominals and an extensive conjugation system for verbs, to name a few. Furthermore, the similarities also extend to the syntax of the languages and orthographies, although with various systematic differences in both. With this in mind it can be supposed that some NLP tools that perform well for Icelandic may also perform well for Faroese, especially data-driven applications.

3 Data collection and preparation

A cornerstone of implementing the ABLTagger system for Faroese is collecting and preparing the resources needed for the task. In the current project, this consisted of standardising the training corpus, revising an already existing tagging scheme for Faroese and compiling an experimental morphological database. This is described in the following section.

3.1 Sosialurin Corpus

The aforementioned Sosialurin project, conducted by Hansen et al. (2004) aimed at gathering text into a sizeable corpus, which would then be machine-tagged and finally manually corrected to create a standardised PoS-tagged corpus for Faroese, to be used in NLP projects and linguistic research. This corpus, hereafter referred to as the Sosialurin corpus, consists of 221 excerpts from the newspaper Sosialurin.[4] In total, the corpus contains 119,833 tokens in 4,073 sentences, about 104,000 words excluding punctuation.

As the corpus was meant to be used in NLP projects and linguistic research in general, the corpus comes prepared in a *token-tag* format, where each line of the corpus file contains a single token and its corresponding PoS tag, separated by a `tab` character. Sentences are demarcated with an empty line and tokens are PoS-tagged using a PoS tagging scheme devised specifically for the corpus.

Before use in the current project, the Sosialurin corpus was slightly modified from the original. This included removing duplicated sentences and metadata from the corpus text and standardising organisation and sentence demarcation. This process produces a revised version of the Sosialurin corpus, bringing the total number of sentences in the revised corpus to 6,156, from the original corpus' 4,073 and the total number of tokens to 117,690 from the original 119,819. It is worth noting this corpus is less than 10% as large as the combined training corpus used to train the original ABLTagger implementation for Icelandic, as discussed in Section 2.1. Nevertheless, as the largest hand-corrected, Faroese PoS tagged corpus, this revised version of the Sosialurin corpus was used to train the ABLTagger implementation for Faroese, with the supposed effect of its relatively small size being discussed further in Section 4.

3.2 Revised Faroese tagging scheme

The tagging scheme devised for the Sosialurin corpus by Hansen et al. (2004) is, to a large extent, based on the tagging scheme used in the IFD corpus for Icelandic (Pind et al., 1991). In the last decades, a number of revisions have been made to the IFD

[4] Accessible at: https://www.sosialurin.fo/.

tagging scheme, mostly to improve tagging efficiency, with the latest version appearing in the most recent release of the MIM-GOLD hand-corrected, PoS-tagged corpus (Barkarson et al., 2020). The same cannot be said about the Sosialurin tagging scheme, as no substantial revisions have been made to it since its inception. As such, before being used to train the ABLTagger system, a number of revisions were applied to the tagging scheme and subsequently to the PoS tags in the Sosialurin corpus itself.

Most of the revisions applied to the Faroese tagging scheme were based on revisions previously applied to the IFD tagging scheme for Icelandic. These include reworked numeral and punctuation tag strings, simplified case governance tagging for adverbs and the removal of a dedicated tag for past participles. Furthermore, various new tag strings were introduced, also based on the IFD tagging scheme, e.g., distinction between different categories of pronouns.

One example of a language-specific revision made on the tagset was the removal of distinction between person (1st, 2nd or 3rd) from verb tags in the original tagset. This was likely a carry-over from the IFD tagging scheme, as Icelandic verbs are morphologically distinct between person in both singular and plural. In Faroese, verbal person is not morphologically apparent in plural forms, and thus should not be relevant to the tagging scheme, in theory. The effect of this revision to the tagging scheme on tagging accuracy is discussed in Section 4.

When applied to the Sosialurin corpus, the total number of unique PoS tags in the corpus was reduced from 390 to 371. This total does not reflect the tagset changes at large, as not all possible tag strings in the tagging scheme are represented in such a small corpus and while many possible tag strings were removed from the tagset, a number of possible tag strings were added as well.

3.3 Experimental Database of Faroese Morphology (EDFM)

The *ABLTagger* system uses a morphological database, a detailed description of the inflection of a language, in tandem with a bi-LSTM-based neural tagger to enhance the accuracy of the tagger as a whole. In the original experiment for Icelandic, the Database of Icelandic Morphology, *DIM* (Bjarnadóttir, 2012; Bjarnadóttir et al., 2019) was used for this purpose, as Icelandic was the language being tagged. As discussed in Section 2.1, *DIM* is an extensive database with hundreds of thousands of unique lemmas and millions of inflected word forms on file, and was applied in the so-called DIM basic format[5] for the project.

No such database, comparable in size, scope or accessibility to *DIM*, exists for Faroese. This situation of course poses a problem when trying to replicate the *ABLTagger* experiment for Faroese, as the morphological component is essential the high accuracy of the tagging mechanism. For the current project, we overcame this by gathering all freely accessible information on Faroese inflection into an Experimental Database of Faroese Morphology (EDFM), formatted in the DIM basic format. This database can then used for Faroese, the same way *DIM* is used in the *ABLTagger* system. As discussed in Section 2.3, Faroese and Icelandic are to large extent morphologically similar, so this approach is at least theoretically applicable.

3.4 Sourcing morphological data

The inflectional data used to build the EDFM was collected from several main sources. In Table 2, the number of paradigms extracted from each data set is shown, also categorised by lexical category.

Faroese Dictionary Database: The largest single morphological description of Faroese is provided by the Dictionary of Faroese (*Føroysk Orðabók*, Poulsen et al. 1997), hereafter referred to as the *OBG* database. This database is accessible and searchable on the website of Sprotin,[6] a Faroese publishing house which provides digital access to various Faroese dictionaries. This database contains 67,488 word entries, of which 65,062 contain inflectional information and were used in the EDFM. Out of these, 5,374 entries (mainly verbs and numerals) contained partial inflectional paradigms that were extended using Python scripts written for the current project.

Faroese naming committee: Along with the OBG database, the Sprotin website hosts a complete list of approved given names in Faroese, with each name's inflection included. Faroese naming laws dictate that only given names that are approved by a governmental naming committee

[5] See: https://bin.arnastofnun.is/DMII/LTdata/s-format/

[6] Accessible at: www.sprotin.fo/dictionaries.

Category	OBG	Names	Wiktionary	Generated	Manual	Total
Adjectives	11,907	-	16	-	-	11,923
Adverbs	1,289	-	-	-	-	1,289
Conjunctions	-	-	-	-	61	61
Interjections	-	-	-	-	115	115
Nouns	46,492	1,667	113	-	-	48,272
Numerals	-	-	-	47	57	104
Prepositions	-	-	-	-	62	62
Pronouns	-	-	-	-	20	20
Verbs	-	-	7	5,327	-	5,334
Total	59,688	1,667	136	5,374	315	**67,180**

Table 2: Contents of the EDFM by lexical category and source database.

can be used officially (Faroese Naming Committee, 2020). This list provided 1,667 Fareoese given names to the EDFM, containing 880 masculine given names and 787 feminine.

Wiktionary data: The English-language Wiktionary contains entries for various Faroese words, most of which contain morphological information. These entries the were accessed via morphological data from the UniMorph project (Kirov et al., 2018; McCarthy et al., 2020), which was originally extracted and generated from Wiktionary data dumps, specifically from June 20, 2015 (Kirov et al., 2016) and is freely available online.[7] Although 3,077 in total, 2,687 of the UniMorph entries were already represented in the entries extracted from the OBG database. Nevertheless, 390 new entries were extracted, further improving the EDFM.

Manual paradigms: A total of 315 entries in the EDFM were manually defined using morphological descriptions of Faroese as guidelines, e.g. Þráinsson et al. (2004). These were mostly pronouns and functional words, in addition to a number of uninflectable words.

3.5 Formatting the EDFM

As discussed in Section 2.1, the morphological component of the original *ABLTagger* experiment for Icelandic was applied in the DIM basic format. As such, the Faroese inflectional data had to be standardised in a similar manner, in order to be applied in the *ABLTagger* system. This was achieved automatically with purpose built scripts.

To illustrate the output format of the EDFM, the entry for the Faroese word *grunnur* 'foundation' in the EDFM is shown in Figure 1. The format consists of lines of comma-seperated fields, with each line containing an inflectional form of a given word. These lines are grouped into word entries, identifieable by the **lemma**, here `grunnur`, as shown in the first field, and the **database ID number**, here `18433`, as shown in the second field. The remaining fields are **lexical category**, here `kk` for masculine noun, **source data set**, here `obg` for the OBG database, **inflected form** of the word and finally the **database specific grammatical tag**, which encodes morphological information on the word form.

```
grunnur;18433;kk;obg;grunnur;NFET
grunnur;18433;kk;obg;grunn;ÞFET
grunnur;18433;kk;obg;grunni;ÞGFET
grunnur;18433;kk;obg;gruns;EFET
grunnur;18433;kk;obg;grunnar;NFFT
grunnur;18433;kk;obg;grunnar;ÞFFT
grunnur;18433;kk;obg;grunnum;ÞGFFT
grunnur;18433;kk;obg;grunna;EFFT
grunnur;18433;kk;obg;grunnurin;NFETgr
grunnur;18433;kk;obg;grunnin;ÞFETgr
grunnur;18433;kk;obg;grunninum;ÞGFETgr
grunnur;18433;kk;obg;grunsins;EFETgr
grunnur;18433;kk;obg;grunnarnir;NFFTgr
grunnur;18433;kk;obg;grunnarnar;ÞFFTgr
grunnur;18433;kk;obg;grunnunum;ÞGFFTgr
grunnur;18433;kk;obg;grunnanna;EFFTgr
```

Figure 1: Excerpt from EDFM output in the DIM basic format: Inflections of *grunnur*.

Ideally, each word entry contains the full inflectional paradigm for the given word. In its current experimental form, the EDFM contains mostly full paradigms, although it contains partial paradigms for certain lexical categories. Furthermore, as the database has not been formally proofread, there are bound to be some errors in the data, especially in the automatically generate paradigms. As will be

[7] Accessible at: https://github.com/unimorph/fao.

discussed in Section 4, this does not disqualify the EDFM from use in LT implementations, such as ABLTagger, although revisions remain a focus for further work on the database.

4 Evaluation of tagger

The *ABLTagger* system as described by (Steingrímsson et al., 2019) uses DyNet[8] (Neubig et al., 2017). The architecture and model hyperparameters used in the evaluation were all unchanged from the original *ABLTagger* experiment.[9] This included stochastic gradient descent training with initial learning rate of 0.13, which decays 5% per epoch, running for 30 epochs for the full model. The hidden layer of the network has 32 layers. The input text, PoS tags and morphological data are vectorized before use in the system and the resulting embeddings for words, characters and the morphological component have 128, 20 and 61 dimensions respectively.

In accordance with previous experiments (see, e.g., Loftsson 2006; Barkarson 2018; Ingólfsdóttir et al. 2019; Steingrímsson et al. 2019), training and evaluation was done via 10-fold cross validation. In this approach, the whole data set at hand is used for both training and testing. This is especially useful when the data set is not large enough to effectively split into dedicated training and testing sets.

4.1 Evaluation setup

Three variables were taken into account to evaluate the Faroese ABLTagger implementation. These were the effect of the tagset revisions discussed in Section 3.2, the effects of the size and contents of the training corpus, and the effect of adding the morphological component described in Section 3.3 to the ABLTagger system.

Tagset revisions: Three models were trained, each using the Sosialurin corpus with a specific tagset with varying amounts of revisions:

- **S-Baseline:** The original Faroese tagset by Hansen et al. (2004)
- **S-Revised-V:** The revised Faroese tagset, with unchanged verbal plural tags (see discussion in Section 3.2)
- **S-Revised:** The fully revised Faroese tagset

[8]The Dynamic Neural Network Toolkit, see http://dynet.io.

[9]The *ABLTagger* source code is available at https://github.com/steinst/ABLTagger.

Corpus size and contents: To evaluate the performance of ABLTagger on small corpora in general, three comparison models were trained, using subsets of the hand-corrected *Icelandic* MIM-GOLD corpus, with each subset corpus being of a relatively similar size and text genre as the Sosialurin corpus, i.e. news articles:

- **MIM-F:** Texts from the newspaper *Frét-tablaðið*. **Size:** 94,224 tokens
- **MIM-M:** Texts from the newspaper *Morgunblaðið*. **Size:** 243,346 tokens
- **MIM-MR:** Texts from the newspaper *Morgunblaðið*, resized to same size as Sosialurin. **Size:** 117,957 tokens

Addition of morphological data: A Faroese model trained on Sosialurin, with the tagset which provies the best overall accuracy, along with the EDFM as the morphological component. Additionally, three reference models were trained using the same Icelandic corpora as above, with the DIM as the morphological component:

- **S-Morph:** Sosialurin with optimal tagset + *EDFM* as morphological component
- **MIM-F-Morph:** MIM-F model + *DIM* as morphological component
- **MIM-M-Morph:** MIM-M model + *DIM* as morphological component
- **MIM-MR-Morph:** MIM-MR model + *DIM* as morphological component

4.2 Results

The evaluation results of the three "tagset models", S-Baseline, S-Revised-V and S-Revised, are is shown in Table 3.

Model	Token	Sentence	Known	Unknown
S-Baseline	88.92%	23.50%	91.70%	55.85%
S-Revised-V	89.75%	24.09%	92.63%	55.76%
S-Revised	**90.12%**	**25.55%**	**93.01%**	**56.01%**

Table 3: Accuracy of taggers trained on different tagsets.

It is apparent that the baseline model trained on the original, unrevised tagging scheme achieved the lowest overall accuracy of the three. Adding only the revisions based on the Icelandic MIM-GOLD corpus to the tagset (S-Revised-V), as described in Section 3.2, resulted in an accuracy gain of 0.83%,

	S-Baseline		S-Revised-V		S-Revised	
No.	Proposed tag > correct tag	Error rate	Proposed tag > correct tag	Error rate	Proposed tag > correct tag	Error rate
1.	ED > EA	2.24%	DN > DG	3.07%	DN > DG	3.35%
2.	EA > ED	1.91%	DG > DN	3.05%	DG > DN	3.17%
3.	VNPP3 > VI	1.59%	VI > VNPP3	1.68%	VNPP > VI	2.15%
4.	VI > VNPP3	1.56%	C > CI	1.65%	VI > VNPP	1.84%
5.	C > CI	1.51%	VNPP3 > VI	1.52%	C > CI	1.69%
6.	EA > EN	1.33%	C > CR	1.33%	CI > C	1.20%
7.	EN > EA	1.16%	CI > C	1.25%	CR > C	1.19%
8.	CI > C	1.16%	CR > C	1.15%	C > CR	1.17%
9.	CR > C	1.13%	DN > C	0.83%	DN > C	0.92%
10.	C > CR	0.98%	C > DN	0.72%	C > DN	0.77%

Table 4: 10 most common errors in tagset evaluation, divided by tagging scheme

equivalent to a total error reduction of 7.51%. Further omission of grammatical person in plural verb PoS tags raised the overall accuracy of the model by another 0.23%, pushing the total error reduction to 10.05% compared to the baseline. The fully revised tagging scheme also achieved the highest scores in whole-sentence accuracy and for both known and unknown tag accuracy, without much loss of grammatical information in the tagging scheme. As such, the S-Revised model was used for the further evaluation steps.

Although the fully revised tagging scheme produces the most accurate model out of the three, there are some systematic errors in the tagging that the revisions do not affect. In Table 4, the 10 most common errors in the models are shown. The most common errors in all the models concern adverbs and prepositions, specifically concerning case governance. In Table 4, these are lines 1.-2. for the revised models and 1.-2. and 6.-7. for the baseline model. These errors play a bigger role in the S-Baseline model, as the PoS tags **EN**, **EA**, **ED**, **EG** refer to prepositions that govern *nominative*, *accusative*, *dative* and *genitive* case, respectively, with substantial ambiguity between tokens that receive these tags. In the revised tagset, these are replaced by **DN** and **DG**, for prepositions[10] that *do not govern case* and those which *do govern case*, respectively, eliminating some of this ambiguity. Similar to these are the various errors concerning conjunctions (or complementisers), i.e., the tag strings starting with **C**. This is mainly caused by the words *at*, *ið* and *sum* (all meaning 'that') which can be variously tagged as conjunctions, relative conjunctions or (in the case of *at*) as infinitive markers (in which case it means 'to').

After removing grammatical person from tags of plural verbs, these tags continue to cause errors. However, the errors in question, underlined in Table 4, are not caused by ambiguity within the plural verb tags themselves. These errors are caused by the Faroese verbal infinitive form, which should receive the tag **VI**, being lexically identical to the active present plural form, which should receive the tag **VNPP** after the tagset revisions. Although at first glance, the table may suggest that this type of error has a higher rate of occurrence in the fully revised model (from 1.68% and 1.52% to 2.34% and 1.86%) this is not the case. This is simply because the revised tagset merges the plural tags into **VNPP**, thus "collecting" the errors of this type. At any rate, the omission of grammatical person in plural verb PoS tags raises the overall accuracy of the model by a substantial amount without much loss of grammatical information.

The three models trained on *MIM-GOLD* subcorpora described above were evaluated in the same manner as the tagset models, with the results shown in Table 5. Also shown in the table are the full baseline model (without morphological data) from the original *ABLtagger* experiment for Icelandic and the S-Revised model, the model which achieved the highest accuracy above, along with the total token count of all the training corpora used.

There seems to be a correlation between corpus size and tagging accuracy; the larger the training corpus, the higher the achieved tagging accu-

[10] In the original tagging scheme prepositions receive a tag string starting with **E**. These are merged with adverbs (**D**) in the revised tagging schemes.

Model	Overall	Known	Unknown	Corp. sz.
S-Revised	90.12%	93.01%	56.01%	117,690
MIM-F	87.28%	93.46%	56.75%	94,224
MIM-MR	88.31%	93.22%	59.78%	117,957
MIM-M	91.03%	94.41%	64.85%	243,346
ABLTagger	93.25%	95.19%	66.84%	590,279

Table 5: Revised Sosialurin model, Icelandic reference models and the original ABLTagger implementation (Steingrímsson et al., 2019) *without morphological data*.

racy is. With this in mind, at 90.06%, the Faroese S-Revised model achieves a relatively high accuracy, surpassing the overall accuracy of two of the smaller Icelandic reference corpora, although not approaching the 93.25% of the original *ABLTagger* baseline model. These accuracy scores are not directly comparable, as although the technical aspects of the models are identical (and, in theory, the text genre of the training corpora), the tagset used for Faroese is still simpler than the one for Icelandic, which may affect the final tagging accuracy. The results do however show that while the accuracy of the Faroese S-Revised model is not comparable to the original *ABLTagger* baseline, it is in the same ballpark as the Icelandic reference models.

The evaluation of the *ABLTagger* model supplemented with EDFM is shown in Table 6.

Model	Token	Sentence	Known	Unknown
S-Revised	90.12%	25.55%	**93.01%**	**56.01%**
S-Morph	**91.40%**	**29.01%**	92.89%	51.41%

Table 6: Sosialurin morphology evaluation results.

The full model with morphological data achieves a overall accuracy of 91.40%, the highest for all the Faroese models. When compared to the S-Revised model, which achieved a 90.12% accuracy, this shows that applying the EDFM raises the final accuracy by 1.28%, amounting to a total error reduction of 12.96%.

The comparison of the S-Morph model to the Icelandic reference models is shown in Table 7. Each model's accuracy is shown along with the accuracy gain provided by the morphological data. In comparison to the Icelandic reference models, the accuracy gain that the EDFM provides when tagging Faroese is comparatively low. Indeed, the DIM contains a much more thorough description of Icelandic than the EDFM does of Faroese and thus should in theory provide better results when applied with *ABLTagger*. However, these results show that despite its experimental nature, the EDFM does indeed serve its purpose in raising the overall tagging accuracy of the ABLTagger system.

Model	Accuracy	Morph. gain
S-Morph	91.40%	+1.28%
MIM-F-Morph	90.69%	+3.41%
MIM-MR-Morph	91.85%	+3.54%
MIM-M-Morph	92.36%	+1.33%

Table 7: Morphology model results and accuracy gain from morphological data.

5 Application and further work

It remains to be discussed how effectively the tagger produced in the current project can be applied in PoS-tagging Faroese text in general. In Table 8, the current project's S-Morph model, hereafter referred to as the Faroese *ABLTagger* model, is compared to the the last dedicated PoS-tagging implementation for Faroese, by Hansen et al. (2004), as mentioned in Section 2.2.[11] Although this tagger used the unrevised Faroese tagset, as discussed in Sections 3.1 and 3.2, and the exact evaluation procedure used is not known, it can serve as a tentative comparison for the current project, in lieu of a previous state-of-the-art tagging implementation for Faroese.

Implementation	Overall	Known	Unknown
Hansen et al. (2004)	87.00%	91.00%	64.70%
Faroese *ABLTagger* model	91.40%	92.89%	51.41%

Table 8: Tagging accuracy for the current project compared to previous best

As is apparent in Table 8, with an overall tagging accuracy of 91.40%, the model produced in the current project returns a substantial improvement on the previous tagger, which achieved an accuracy of 87.00%. The result achieved by the Faroese *ABLTagger* model is quite promising, especially as it is uses a quite fine-grained tagset. By these metrics, the current project has produced the most accurate dedicated, fine-grained PoS tagger for Faroese to date.

[11] The Faroese Giellatekno implementations (Trosterud, 2009), although not containing a PoS tagger per se, do contain a rule based grammatical analyser, which can function somewhat like a PoS tagger. However, these implementations have not been evaluated in a similar way to the taggers discussed here and are thus left out of the discussion. The possibility of future comparisons remains.

Despite the high reported accuracy of the Faroese model, two issues must be kept in mind. Firstly, the overall accuracy, although high, does not approach the accuracy of the full *ABLTagger* model for Icelandic, which, as discussed in Section 2.3, has a similar tagset and overall morphology to Faroese. In theory, a substantially higher tagging accuracy should be obtainable for Faroese with the *ABLTagger* system, but it is limited by the size and contents of the training data used. Secondly, as the Sosialurin corpus only consists of news articles, its contents are likely not representative of Faroese-language texts in general. The Faroese *ABLTagger* model would thus perform well on news-like texts in its current form, but likely return sub-optimal results when tagging large, unseen texts in different genres, which is the main goal when developing a PoS tagger.

6 Conclusion

In this paper we have described the developement of a dedicated, high-accuracy PoS-tagging solution for Faroese. This was achieved by training *ABLTagger*, a state-of-the-art PoS-tagging system for Icelandic, on Faroese language data which had been revised and formatted with methods and tools based on Icelandic NLP research. This produced a Faroese PoS-tagging model which achieves a 91.40% overall tagging accuracy, when trained on the 100,000 word Sosialurin corpus and evaluated using 10-fold cross validation. The last similar PoS-tagging implementation for Faroese achieved a 87.0% overall accuracy. Thus, in the absence of recent comparable implementations, the Faroese *ABLTagger* model may tentatively be considered the state-of-the-art for PoS-tagging Faroese.

In addition to developing the PoS tagger, this project produced various resources which could prove useful in further research in Faroese language technology. These include a proposed revised PoS-tagging scheme for Faroese, mainly based on the Icelandic MIM-gold tagging scheme, as well as the standardised and revised version of the corresponding Sosialurin PoS-tagged Corpus. The same goes for the EDFM, the experimental morphological database compiled for use with the tagger. Although its format is based on its Icelandic counterpart, DIM, and is mostly compiled from already existing Faroese dictionary data, with 67,488 word forms and about 1,000,000 inflectional forms, it is the first of its kind for Faroese as a single, accessible data set designed for use in language technology implementations.

As Faroese digital language resources are, at the moment, few and far between, Faroese language technology has ground to cover before it can be considered fully equipped to tackle recent innovations in the field. As the data sets and PoS-tagging model produced in this project have been made available online,[12] they may well serve as a basis for further developments for Fareoese, both to implement new NLP applications and provide further opportunities for linguistic research.

References

Starkaður Barkarson. 2018. Þjálfun málfræðimarkarans Stagger með nýjum gullstaðli [Training of the PoS tagger Stagger with a New Gold Standard]. Unpublished MA thesis. URL http://hdl.handle.net/1946/29474.

Starkaður Barkarson, Einar Freyr Sigurðsson, Eiríkur Rögnvaldsson, Hildur Hafsteinsdóttir, Hrafn Loftsson, Steinþór Steingrímsson, and Þórdís Dröfn Andrésdóttir. 2020. MIM-GOLD 20.05. CLARIN-IS, Stofnun Árna Magnússonar.

Eckhart Bick, Heini Justinussen, Zakaris Svabo Hansen, Trond Trosterud, and Tino Didriksen. 2020. Corpuseye Faroese Corpus.

Kristín Bjarnadóttir. 2012. The database of modern Icelandic inflection (Beygingarlýsing íslensks nútímamáls). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SaLTMiL 8 – AfLaT2012)*, pages 13–20, Istanbul, Turkey. European Language Resources Association.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. Dim: The database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland. Linköping University Electronic Press.

Thorsten Brants. 2000. TnT: A statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, page 224–231, USA. Association for Computational Linguistics.

Faroese Naming Committee. 2020. Góðkend Fólkanøvn [Approved Given Names].

Zakaris Svabo Hansen, Heini Justinussen, and Mortan Ólason. 2004. Marking av teldutøkum tekstsavni [Tagging of a digital text corpus].

[12] See: https://github.com/hinrikur/far-ABLTagger.

Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Joel Wallenberg. 2014. Rapid Deployment of Phrase Structure Parsing for Related languages: A Case Study of Insular Scandinavian. In *Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 91–95. European Language Resources Association (ELRA).

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

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3121–3126. European Language Resources Association (ELRA).

Hrafn Loftsson. 2006. Tagging Icelandic text: An experiment with integrations and combinations of taggers. *Language Resources and Evaluation*, 40(2):175–181.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. Unimorph 3.0: Universal morphology. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin-Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Ku-mar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *CoRR, abs/1701.03980*.

Anna Björk Nikulásdóttir, Jón Guðnason, and Steinþór Steingrímsson. 2017. *Mál tækni fyrir íslensku 2018–2022: verkáætlun [Language Technology for Icelandic 2018-2022: Strategic Plan]*. Mennta- og menningarmálaráðuneytið, Reykjavík, Iceland.

Jörgen Pind, Friðrik Magnússon, and Stefán Briem. 1991. *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavík, Iceland.

Jóhan Hendrik W. Poulsen, Marjun Simonsen, Jógvan í Lón Jacobsen, Anfinnur Jóhansen, and Zakaris Svabo Hansen, editors. 1997. *Føroysk orðabók [Dictionary of Faroese]*. Føroya fróðskaparfelag, Torshavn.

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 the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Bulgaria.

Trond Trosterud. 2009. A constraint grammar for Faroese. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2009)*. Northern European Association for Language Technology (NEALT).

Höskuldur Þráinsson, Hjalmar P. Petersen, Jógvan í Lón Jacobsen, and Zakaris Svabo Hansen. 2004. *Faroese: An overview and reference grammar*. Føroya fróðskaparfelag, Torshavn.