To ð or not to ð
A Faroese CG-based grammar checker targeting ð errors

Trond Trosterud
UiT The Arctic University of Norway
trond.trosterud@uit.no

Abstract

Many errors in Faroese writing are linked to the letter *ð*, a letter which has no corresponding phoneme, and is always omitted intervocally and wordfinally after a vowel. It plays an important role in the written language, disambiguating homophone but not homograph forms like infinitive *kasta* 'throw' from its participle *kastað*. Since adding a hypercorrect *ð* or erroneously omitting it often results in an existing word, these errors cannot be captured by ordinary spellcheckers. The article presents a grammar checker targeting *ð* errors, and discusses challenges related to false alarms.

1 Introduction

The article addresses a central problem in written Faroese: How to correct errors arising from erroneously writing or omitting the letter *ð* in such a way that the resulting errouneous form is an existing word. A typical case of *ð* omission is (1), and an instance of superfluous *ð* is (2)[1].

(1) Tey hava serliga ***tosa** um at í
they have especially talk.V.Inf about that in
Grønlandi er tað grønlendskt sum skal vera
Greenland is it Greenlandic that shall be
fyrsta mál.
first language
'In particular, they have talked about the fact that in Greenland Greenlandic shall be official language'

(2) Eg ***haldið** at orsøkin til at HB
I consider.V.Imp.Pl that reason for that HB
vann móti KÍ var ein einastandandi
won against KÍ was a unique
liðinnsatsur.
team.effort
'In my opinion, the reason why HB won against KÍ was an outstanding effort by the team'

In (1), *ð* is omitted from the correct supine form *tosað*, resulting in an infinitive, and in (2) the correct present first person singular form *haldi* has received a hypercorrect *ð*, resulting in a plural imperative form.

The challenge is to correct such errors. The approach presented here is to build a grammar checker on top of a grammatical analysis of Faroese, where the erroneous patterns are identified and the correct forms are presented to the user, accompanied by an explanation. The grammar checker is already part of the web-based version of the Faroese spell checker[2], and the main challenge at the present stage is thus to have a good precision. Testing the recall of the grammar checker is obviously relevant for a thorough evaluation, but this falls outside the scope of the present article.

The article is structured as follows. First, section 2 shortly presents relevant aspects of Faroese and of the morphological and grammatical components providing the input to the grammar checker. Section 3 presents the grammar checker. Section 4 presents the evaluation material and discusses the results. Finally comes a conclusion.

2 Background

2.1 Faroese grammar and the letter *ð*

Faroese is a North Germanic language spoken by appr. 80.000 people, mainly on the Faroe Islands. The grammatical structure of written Faroese contains the traditional three gender (masculine, feminine, neuter) and four case (nominative, accusative, genitive, dative) system and person inflection for verbs known from Old Norse and Ice-

[1] In the examples, the wordform flagged as an error will be given in **bold**. When the wordform is wrong, it is marked with an asterisk. When it is correct, and the alarm is false, there is no asterisk.

[2] https://divvun.no/korrektur/gramcheck.html

Proceedings of the NoDaLiDa 2023 Workshop on Constraint Grammar - Methods, Tools and Applications, pages 15–19

landic. Contrary to these languages, person inflection in Faroese is found only in the singular. For a presentation, see (Thráinsson et al., 2012).

Faroese orthography is conservative and the written standard differs considerably from the spoken language, which itself is divided in several dialects. Relevant to the present discussion is the letter *ð*, which plays a central role in the inflectional system of the written language. The *ð* may be added to both nominal stems, giving definite forms, and verbal stems, giving participles or imperative plural forms. As shown by (Thráinsson et al., 2012) p. 20, "the letter *ð* [does] not as a rule have any phonetic value intervocalically or word-finally after a vowel". Word form pairs distinguished by the *-ð* suffix thus give rise to homonymy pairs in speech, but not in writing. Central homonymy cases are shown in table 1[3].

MS cat.	Form	MS cat.	Form
V.Inf & Prs.Pl.	kalla	Ptc. & Sup.	kallað
V.Prs.Sg1 & N.Dat.Indef.	fari	Ptc & Sup N.Nom.Def N.Acc.Def	farið
V.Inf & A.Def	norska	Ptc. & Sup.	norskað

Table 1: Systematic homonymies. Example words: *kalla* V 'call, name', *fara* V 'leave, travel', *far* N 'track', *norsk* A 'Norwegian', *norska* V 'make Norwegian'.

2.2 Faroese morphology and disambiguation

The Faroese morphology is handled by a finite state transducer (Beesley and Karttunen, 2003), described in (Trosterud, 2009). The morphological description was mainly based upon (Thráinsson et al., 2004), but in order to get a comprehensive description of the morphology, the transducer was built with the inflection classes from (Poulsen et al., 1998). The lexicon was based upon (Poulsen et al., 1998), but complemented with frequent words from the online Faroese corpus[4]. Issues not covered by these sources were addressed in cooperation with Heðin Jákupsson.

Faroese inflectional morphology is rich in homonymy, with on average 4.0 analyses per word form. In order to disambiguate this, the grammar checker uses a disambiguator based upon constraint grammar (Karlsson, 1990). The constraint grammar is presented in (Trosterud, 2009).

3 The Faroese grammar checker

3.1 Technical background

The system is built on a pipeline of modules as presented in Wiechetek (2019). The pipeline uses the free open source implementation HFST (Lindén et al., 2013) for finite-state automata and VISLCG-3 (Didriksen, 2016) for constraint grammar. Both are included in the *GiellaLT* infrastructure (cf. Moshagen et. al., (2013) for a presentation).

The grammar checker uses the finite state transducer presented in 2.2, but instead of the ordinary disambiguator it uses a relaxed version of it. The reason for this is that the disambiguator presented in 2.2 is based upon the assumption that the input is correct. Since this assumption does not hold for a grammar checker, certain disambiguation rules had to be relaxed in order not to remove relevant target forms.

The Faroese grammar checker is part of a multilingual infrastructure *GiellaLT*, which includes language models either released or on a functional (beta) level for appr. 40 languages. The source code is publicly available[5].

The Faroese grammar checker is already available for use in the Divvun grammar checker interface [6]. Given that the grammar checker is still in an early stage, its main purpose is to make the Faroese spell checker (which is integrated in the grammar checker) available also on Google docs and on MS Word for Macintosh, platforms who do not allow third-party spell checkers. For the present stage of the grammar checker development it is thus more important to avoid false alarms than to achieve a good coverage.

3.2 Errors to be targeted

In this article, only a part of the grammar checker rule set is presented, the one relevant to a certain type of *ð* errors, errors due to spoken language homonymy due to *ð* suffixes in one of the forms. The errors targeted are the confusion of supine (= neuter participle when combined with an auxiliary) and infinitive forms, the confusion of participle

[3] Abbreviations: Prs = present tense, Ptc = participle, Sup = supine, Indef/Def = (in)definite

[4] `https://gtweb.uit.no/f_korp`

[5] The source code for Faroese is found here: `https://github.com/giellalt/lang-fao`.

[6] The Divvun grammar checker interface makes it possible to use the grammar checker together with MS Word and Google docs, cf. `https://divvun.no/en/korrektur/gramcheck.html`

and first person singular forms, as well as the confusion of supine and present plural forms.

4 Evaluation

4.1 The material

As evaluation corpus was used a subset of the Faroese BLARK text corpus (Simonsen et al., 2022). It contained 9.0 million words, from the following genres (table 2):

Genre	Words
Students 17-20 years	77.674
Magazines	339.751
Blogs	285.637
Online news	7.180.722
Newspapers	1.138.988
Total	9.022.772

Table 2: Text genres in the test corpus

The largest category is online news, containing texts both from the Faroese Broadcasting company KVF and the online news portal website *Portalurin*. The magazines included are *MEGD* and *Starvsbladid*. More details are given in the metadata of the BLARK itself.

4.2 Results and analysis

The corpus was run through the grammar checker[7], and each alarm (reported error) was manualy evaluated. Looking at the results by genre, we get the results shown in table 3. For each genre, the table gives the number of alarms (cases the grammar checker flags as erroneous) as well as whether they actually are wrong (TP, or true positive) or not (FP, or false positive). Precision is calculated as the number of true positives divided by all alarms.

Genre	**Alrms**	**Alrms /100k**	**TP**	**FP**	**Prec. (%)**
17-20yrs	9	11.6	7	2	77.8
Mags	30	8.8	23	7	76.7
Blogs	20	7.0	11	9	55.0
Onl.nws	370	5.2	274	96	64.7
Newsp.	2	0.2	2	0	100.0
Total	431	4.8	317	114	73.5

Table 3: Evaluation

[7]The grammar checker used for testing was the version from Nov 4th 2022, github.com/giellalt/lang-fao/blob/main/tools/grammarcheckers/grammarchecker.cg3

For all genres the percentage of alarms was low, around or below ten per 100.000 words. As can be seen, the errors are somewhat more common for genres where we would expect less proofreading. Investigating recall is outside the scope of the present paper, but it seems likely that only a part of the real amount of (relevant) errors has been found. Precision, or the percentage of correct alarms, varies from genre to genre, with 73.5 % calculated on the corpus as a whole.

Looking now at the alarms according to grammatical type, we get a different picture, with more variation in the precision. Table 4 gives an overview. The rule types are written on the format *wrong form* → *correct form*.

Rule	**Total**	**TP**	**FP**	**Prec.**
sup → inf	44	37	7	84.1 %
inf → sup	287	230	57	80.1 %
prfptc → sg1	8	6	2	75.0 %
sup → sg1	56	30	28	53.6 %
sup → prspl	36	14	23	38.9 %
Total	431	317	117	73.5 %

Table 4: Alarms according to rule type

The most common error type was infinitive for supine, the type shown in (1). It contained 66.5 % of all the alarms in the evaluation material. The error type also had a good precision rate, 80.1 %.

The false alarms typically involved errors in part of speech disambiguation. A case in point is the false alarm shown in (3).

(3) Eg havi **illgruna** um
I have.V.Prs.Sg1 suspicion.N.Sg.Acc about
at tað er tí mótargument
that.Sbj that.Det is because counter.argument
mangla, ella hvussu?
is.missing, or what?
'My suspicion is that this is because the counter arguments are missing, don't you agree?'

The form *illgruna* is also a verb, with a participle *illgrunað*. The grammar checker has thus erroneously identified it as an infinitive-for-supine pattern. The quite frequent form *illgruna* occurred in several false alarms, and should be identified as part of the collocation *hava illgruna um* 'be suspicious about'.

Another false alarm, this case one of accidental and not systematic homonymy, is (4).

(4) Hava vit ikki **egna**
have.V.Prs.Pl we not own.A.Sg.Acc.Indef
søgu, mál og identitet?
history, language and identity
'Don't we have our own history, language and identity'

Here, the accusative form of the common adjective *egin* 'own' is accidently identical to the verb *egna* 'to bait, to add fishbait on the hook'. In a revised version this should be solved by including *egna* in a set of infinitives not to be corrected. Almost all false alarms for this rule were of these two types.

The inverse error type, supine for (correct) infinitive, shown in (5), was more rare, with 10 % of the alarms. This type showed the best precision of all the error types.

(5) Ja hvat annað skal man ***tosað** um?
Yes what else shall one talk.V.Sup about?
'Well, what else should one have talked about?'

For this rule type, some of the false alarms were due to the pronoun *man* 'one', that (probably for puristic reasons) was not included in the Faroese dictionary (Poulsen et al., 1998) and therefore also not in the language model, and thus was confused for the homonymous present singular form of the modal *munna* 'may'. An example of this type is (6).

(6) Tað sær út til, at øll hesi árini
that looks out to, that all these years
hevur man ikki **megnað** at fáa
have.V.Prs.Sg3 one not achieve.V.Sup to get
broytingar í tær samsýningar, sum eru, sigur
changes in the fees, that are, says
løgmaður
lawyer
'It looks like one during all these years has not been able to get any changes in the existing fees, the lawyer says.'

The two next error types represent hypercorrect use of *ð* in first person singular form, as in example (2) above. Another example is (7).

(7) Eg ***sitið** eitt mjørkatungt
I sit.V.Sup one dark.heavy
summarkvøld í einum hugnaligum
summer.evening in one cosy
køki í Havn
kitchen in Tórshavn
'A dark summer evening I sit in a cosy kitchen in Torshavn'

For this error type, the precision was lower than for the supine/infinitive ones. The main problem for these rules was that they failed to capture a first person verb *havi* 'have.V.Prs.Sg1' to the left (8).

(8) Mangan havi eg **sitið** og verið
Often have I sit.V.Sup and been
ónøgd við, at meira ikki hevur verið
dissatisfied with, at more not has been
gjørt til tess at vinna okkum betri sømdir
done in.order.to get us better regard
'Many a time I have been dissatisfied by the fact that not more has been done in order to achieve a better reputation'

The problem was the preceeding disambiguation rule, which errouneously removed the verb reading of *havi* due to a typo in the tag for first person pronouns. *havi* was then analysed as a noun, and the grammar checker thus flagged *sitið* as an error.

A further weakness of the grammar checker revealed during evaluation was that it flagged Sg1 errors also when the target form did not end in *-ið*.

(9) Í mong harrans ár havi eg **skrivað** til
in many Lord's years have I write.V.Sup to
damubløðini tey kalla, men altíð
women's.magazines they say, but always
undir dulnevni.
under pseudonym
'For God knows how many years I have written to the so-called women's magazines, but always under pseudonym'

The point here is that *skrivað* is not a likely misspelling of first person *skrivi*, contrary to *sitið/siti*. The rule should thus have been restricted to the inflection classes with supine forms in *-ið*.

When it finds a potential error, the grammar checker suggests a form to replace it, whenever possible. In some cases the error identification was correct whereas the suggestion was not. One example is the supine form of *vera* 'to be', which is *verið*. This form occurred in several correctly flagged Sup → Sg1 errors, e.g. (10). .

(10) Eg ***verið** fullkomiliga
I be.V.Sup completely
fríkendur
aquit.V.PrfPtc.Msc.Sg.Nom.Indef
'I was completely acquitted'

Since the rules assume that the confused forms are supine and first person singular, it suggested the form *eri*, the first person present of *vera*. The

form *verið*, however, is not a likely confusion of *eri*. It turned out that the target form here was not the copula, but the verb *verða* 'to become', which first person form is *verði*. Since the *ð* is not pronounced in this phonological context, the form is a homonym of *verið*. What is called for is thus a separate rule for this important verb, suggesting *verði* whenever *verið* occurs in first person singular contexts.

5 Conclusion

This article has presented an early version of a Faroese grammar checker, targeting errors related to inflectional forms containing the suffix *ð*. Even though the grammar checker still contains some obvious errors, the precision is quite good, over 80 % for the most frequent *ð* error type. With these errors corrected as well as an improved suggestion component, the present grammar checker may be seen as both a welcome addition to the Faroese spell checker as well as a pedagogical tool for pupils during the learning process.

The next steps for the grammar checker are to investigate the recall of the error types it already covers (to look at the *ð* errors the grammar checker fails to capture), and to include more error types. This is left for future research.

Acknowledgments

Thanks to my Faroese colleagues at Setur for inspiring discussions, to Heðin Jákupsson for cooperation on improving the morphological model, to Hanna Jensen for help with some tricky sentences in the evaluation dataset as well as for an analysis of the *verið* cases, and to the anonymous reviewers for useful comments.

References

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. Studies in Computational Linguistics. CSLI Publications, Stanford, California.

Tino Didriksen. 2016. *Constraint Grammar Manual: 3rd version of the CG formalism variant*. GrammarSoft ApS, Denmark.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *COLING '90 Proceedings of the 13th conference on Computational linguistics*, volume 3, pages 168–173, Helsinki.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Miikka Silfverberg, and Tommi A. Pirinen. 2013. Using HFST for creating computational linguistic applications. In Adam Przepiórkowski, Maciej Piasecki, Krzysztof Jassem, and Piotr Fuglewicz, editors, *Computational Linguistics: Applications*, pages 3–25. Springer Berlin Heidelberg, Berlin, Heidelberg.

Sjur N. Moshagen, Tommi Pirinen, and Trond Trosterud. 2013. Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22–24; 2013; Oslo University; Norway*, number 16 in NEALT Proceedings Series, pages 343–352. Linköping University Electronic Press.

Jóhan Hendrik W. Poulsen, Marjun Simonsen, Jógvan í Lon Jacobsen, Anfinnur Johansen, and Zacharis Svabo Hansen. 1998. *Føroysk orðabók*, volume 1-2. Føroya Fróðskaparfelag, Tórshavn.

Annika Simonsen, Sandra Saxov Lamhauge, Iben Nyholm Debess, and Peter Juel Henrichsen. 2022. Creating a basic language resource kit for Faroese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4637–4643, Marseille, France. European Language Resources Association.

Höskuldur Thráinsson, Hjalmar P. Petersen, Jógvan í Lon Jacobsen, and Zacharis Svabo Hansen. 2004. *Faroese: An overview and reference grammar*. Føroya Fróðskaparfelag, Tórshavn.

Höskuldur Thráinsson, Hjalmar P. Petersen, Jógvan í Lon Jacobsen, and Zacharis Svabo Hansen. 2012. *Faroese: An overview and reference grammar*. Føroya Fróðskaparfelag, Tórshavn.

Trond Trosterud. 2009. A constraint grammar for Faroese. In *Proceedings of the 17th Nordic Conference of Computational Linguistics. NEALT Proceedings Series*, volume 4, pages 1–7.

Linda Wiechetek, Sjur Moshagen, Børre Gaup, and Thomas Omma. 2019. Many shades of grammar checking – Launching a Constraint Grammar tool for North Sámi. In *Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar: Methods, Tools and Applications, Turku, Finland*, volume 33 of *NEALT Proceedings Series*, Linköping, Sweden. Linköping University Electronic Press.