A constraint grammar for Faroese

Trond Trosterud
University of Tromsø

Abstract

The present paper presents ongoing work on a finite-state transducer, a Constraint Grammar disambiguator and dependency grammar for Faroese. In Faroese, the classical Germanic system of case, person and number inflection is upheld, but with somewhat more homonymy than in the closely related Icelandic. Rather than conflating homonym categories, the present morphological transducer gives a fully specified analysis of all morphological distinctions.

1 Introduction

The transducer is based upon the lemma list of *Føroysk orðabók* ((Poulsen et al., 1998))[1], and upon the grammatical description found in (Thráinsson et al., 2004).

The Faroese parser uses the computational infrastructure from the Sámi parser project (*giellatekno.uit.no*). It has the same file setup, similar makefiles, etc. There are also benefits in the opposite relation: The Sámi morphophonology testsuite was taken from work on the Faroese *twolc* file.

The Faroese morphological analyser/generator *Ffst* is a finite-state transducer. It is compiled with Xerox transducer compilers: *twolc* for Morphophonology, and *lexc* for lexicon and morphology (cf. (Beesley and Karttunen, 2003) and `http://www.fsmbook.com/`). The disambiguator *Fdis* and dependency grammar *Fdep* are written within the Constraint Grammar framework (see e.g. (Karlsson, 1990), (Karlsson et al., 1995)), and uses the 3rd generation compiler *vislcg3* ((Bick, 2000), `http://beta.visl.sdu.dk/cg3.html`).

[1]Thanks to the authors for making lemmalist and inflection codes electronically accessible, Without it this project would of course not have been realisable.

Sections 2, 3 and 4 present the grammatical analyser Ffst, the disambiguator Fdis and the dependency grammar Fdep, respectively. Section 5 gives an evaluation of the current stand of the parser, and the final section contains future perspectives and a conclusion.

2 The grammatical analyser

2.1 Lexicon

The Ffst lexicon uses the same inflectional codes as does (Poulsen et al., 1998). Dictionary updates and new words annotated with the same codes may thus be added directly to the Ffst. The analyser has a dynamic compounding component, genitive singular nouns have the basic noun lexicon as one of their continuation lexica, thereby creating a loop allowing any compound with genitive singular first part. This gives rise to a circular transducer, for generation this component must thus be switched off.

Ffst also contains a name guesser. The guesser detects words with capital first letter and non-Faroese phonotax. The candidate words must contain at least one vowel. The final letter cannot be a Faroese suffixal sound (*a, i, u, n, m, r, s, t* (to avoid explicit case endings). The putative name is then assigned Nom, Acc and Dat. If there is any other analysis available, the guessed form is automatically discarded. The guesser is very reliable: Of the 500 most common guesses all 500 were actually names. It is also (too) careful: Banning Faroese case suffixes from the guesser avoids analysig case-inflected forms as baseforms, but at the same time it prevents the parser from making many correct guesses.

2.2 Morphology

The morphological part of *Fdis* is built in several layers. For the nominal morphology, the first layer gives the part of speech and gender tags, and mor-

phophonological flags, as shown in Figure 1, for the noun *bóndi* "farmer", where the nominative and accusative plural forms show Umlaut.

```
LEXICON k5 ! bóndi
+N+Msc:          W-M-SGNOM ;
+N+Msc:          W-M-SGACC ;
+N+Msc:          W-M-SGDAT ;
+N+Msc:          W-M-SGGEN ;
+N+Msc:%^IUML    W-M-PLNOM-UR ;
+N+Msc:%^IUML    W-M-PLACC-UR ;
+N+Msc:          W-M-PLDAT ;
+N+Msc:          W-M-PLGEN ;
```

Figure 1: Definiteness morphology

The dictionary contains xx nominal declension types, but including singular-only and plural-only declension patterns, and combined patterns (words declined for more than one pattern), the system totals 269 distinct first-layer continuation lexica for nouns, one of them being the $k5$ lexicon in Figure 1.

The second layer gives case and number morphology. Figure 2 gives the continuation lexica for weak masculine plurals, i.e., also for *bóndi* and the other $k5$ words.

```
! Plural

LEXICON W-M-PLNOM
+Pl+Nom:%>ar DF-N-PLm ;

LEXICON W-M-PLNOM-UR
+Pl+Nom:%>ur DF-N-PLm ;

LEXICON W-M-PLACC
+Pl+Acc:%>ar DF-A-PLm ;

LEXICON W-M-PLACC-UR
+Pl+Acc:%>ur DF-A-PLm ;

LEXICON W-M-PLDAT
+Pl+Dat:%>u DF-D-PL ;

LEXICON W-M-PLGEN
+Pl+Gen:%>a DF-G-PL ;
```

Figure 2: Second layer - case and number

The third layer gives definiteness morphology. Due to the agglutinative nature of Faroese morphology, the lexica either only add the indefinite tag, or the definite tag and suffix. The exception is dative, which shows an *n:m* alternation. Rather than writing a morphophonological rule deleting *m* in front of *num*, the alternation is written into the morphology file.

```
LEXICON DF-N-PLm
+Indef:    # ;
+Def:%>nir # ;

LEXICON DF-A-PLm
+Indef:    # ;
+Def:%>nar # ;

LEXICON DF-D-PL
+Indef:%>m # ;
+Def:%>num # ;

LEXICON DF-G-PL
+Indef:  # ;
+Def:nna # ;
```

Figure 3: Third layer - definiteness

Applying these lexica, we get, among others the accusative and dative plural definite forms shown in Figure 4.

```
bóndi+N+Pl+Acc+Def
bónd%>^IUML%>ur%>nar

bóndi+N+Pl+Dat+Def
bónd%>u%>num
```

Figure 4: The resulting upper and lower lexc strings

2.3 Morphophonology

The lower part of the string pairs from the morphological transducer are then fed to a separate automaton, the morphophonological component. This automaton contains rules for morphophonological alternations, and for non-segmental morphology. The relevant rule in this context is *I-umlaut*, shown in figure 5. The rule works on strings containing any of the vowels in Vx, zero or more consonants, and the Umlaut trigger symbol ^IUML, and changes all vowels in Vx into the corresponding vowels in Vy. In this case, it changes *ó* into *ø*.

```
"I-umlaut"
 Vx:Vy <=> _ Cns* %^IUML: ;
     where Vx in ( a o ø á ó ú )
           Vy in ( e y i æ ø ý )
       matched ;
```

Figure 5: The twolc i-umlaut rule

The morphological and morphophonological transducers are then composed, and the resulting transducers gives a pairing of the upper represenation of the former and the lower represntation of the latter, graphically presented in Figure 6, with the invisible, intermediate strings shown in shaded grey.

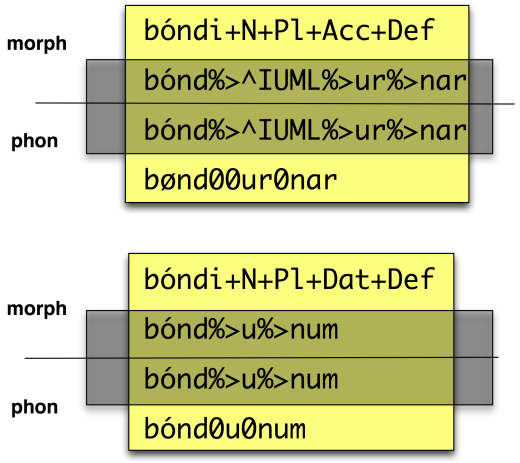

Figure 6: The transducers

Applied to all grammatical words of the lexeme *bóndi*, Ffst gives the paradigm shown in Figure 7.

bóndi+N+Msc+Sg+Nom+Indef	bóndi
bóndi+N+Msc+Sg+Acc+Indef	bónda
bóndi+N+Msc+Sg+Dat+Indef	bónda
bóndi+N+Msc+Sg+Gen+Indef	bónda
bóndi+N+Msc+Pl+Nom+Indef	bøndur
bóndi+N+Msc+Pl+Acc+Indef	bøndur
bóndi+N+Msc+Pl+Dat+Indef	bóndum
bóndi+N+Msc+Pl+Gen+Indef	bónda
bóndi+N+Msc+Sg+Nom+Def	bóndin
bóndi+N+Msc+Sg+Acc+Def	bóndan
bóndi+N+Msc+Sg+Dat+Def	bóndanum
bóndi+N+Msc+Sg+Gen+Def	bóndans
bóndi+N+Msc+Pl+Nom+Def	bøndurnir
bóndi+N+Msc+Pl+Acc+Def	bøndurnar
bóndi+N+Msc+Pl+Dat+Def	bóndunum
bóndi+N+Msc+Pl+Gen+Def	bóndanna

Figure 7: The resulting paradigm for bóndi

A list of the morphophonological rules is given in Figure 8 on page 4.

2.4 Status quo for Ffst

At present (May 2009), the Faroese morphological transducer recognises 94.3 % of all wordform tokens and 62.8 % of all wordform types in running text, for a corpus of 2.7 million words (dominated by newspaper text). The discrepancy indicates that Ffst handles common words better than rare ones.

The results could still be better, but for certain subgenres (such as the Bible), Ffst gives better results (96.3 % and 83.3 %, respectively), results good enough to evaluate the subsequent CG component. Note that even for the known text, Ffst misses approximately 16 % of the wordform types. The reason for this high number is that certain parts of the transducer are still under construction, especially parts of the irregular verbs, and of comparative and superlative forms of adjectives. Also Faroese names are missing, except the most central person names. The foreign names are mainly taken care of by the name guesser.

The top 84 missing wordforms from an 2.7 m wd corpus are shown in Figure 9 on page 4.

The 43093 missing wordforms represent 5.67% of the 2.7 mill corpus. In order to reduce the number of missing wordforms in running text by 50%, the top 2117 wordforms of the missing list would have to be added to the analyser. Important areas for lexicon improvement include the following:

- Adjectival inflection of participles, irregular adjectival forms
- Some irregular strong verbs and verb forms
- Faroese names (other than person names)
- Compounded function words
- Words missing from FO
- Plain errors

3 The Faroese disambiguator

The disambiguator (Fdis) consists of 166 rules for morphological disambiguation, 67 mapping rules, and 68 rules for disambiguation of grammatical functions. This is a small, but relatively efficient rule set, compared to the disambiguators for some other languages in Table 1[2]. For each language, the table gives number or rules, and the average numbers of readings before and after disambiguation, as applied on a compatible corpus (Genesis and the New Testament.).

3.1 Tag unification

The efficiency of the Fdis ruleset illustrates the efficiency of an innovation in vislcg3, namely set

[2]The Sámi parsers are developed at the University of Tromsø (UiT), the Greenlandic parser is joint work between Oqaasileriffik and UiT, and the Bokmål parser is developed at Tesktlaboratoriet in Oslo. Thanks to Kristin Hagen for running the Bokmål analysis for this comparison.

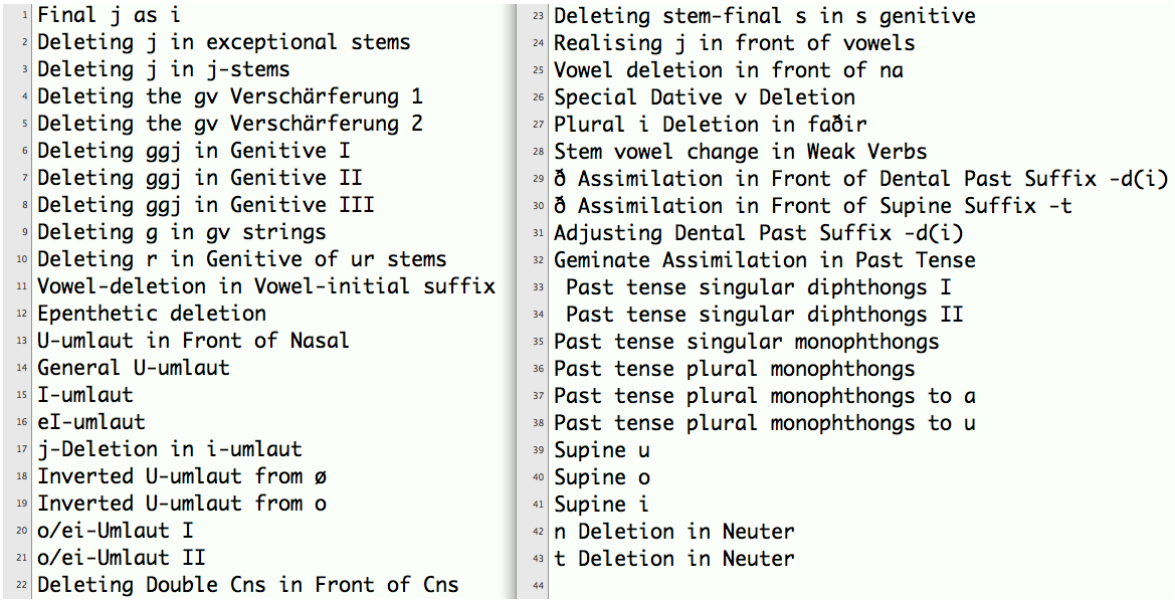

Figure 8: Twol rules

5,67%	komnir	5,45%	fyrrenn	5,29%	irakska
5,66%	samtykti	5,44%	Almanna-	5,29%	hesaferð
5,65%	tykist	5,44%	The	5,28%	Reynatúgvu
5,64%	eydnaðist	5,43%	Efra	5,28%	geri
5,63%	egnu	5,42%	tiknir	5,27%	Norðurlandaráðnum
5,62%	mist	5,42%	finst	5,27%	tiknar
5,61%	farnir	5,41%	Klæmint	5,26%	komnar
5,60%	nærum	5,40%	kravdi	5,26%	r
5,59%	gingið	5,40%	Himli	5,25%	liðnum
5,58%	selt	5,39%	Skipara-	5,25%	krígnum
5,57%	a.	5,39%	Fiskimálaráðnum	5,24%	Bjarkhamar
5,57%	snýr	5,38%	nærri	5,24%	toppskjútti
5,56%	miljónir	5,37%	virksemið	5,23%	misti
5,55%	doyðu	5,37%	himli	5,23%	hálvgum
5,54%	Vinnumálaráðið	5,36%	Eyðun	5,22%	tinglimir
5,53%	Maskinmeistarafelagið	5,36%	yvirgangi	5,22%	skuldsettur
5,53%	íløgur	5,35%	høgru	5,21%	Innlendismálaráðið
5,52%	fiskiskap	5,34%	Norðoyatunnilin	5,21%	umhugsa
5,51%	oynni	5,34%	forsetavalið	5,20%	rundanum
5,50%	elstu	5,33%	samfelagnum	5,20%	forkvinna
5,50%	æt	5,33%	gomul	5,19%	økjum
5,49%	býr	5,32%	blaðnum	5,19%	innanhýsis
5,48%	herurin	5,32%	uppat	5,18%	Ba
5,48%	farnu	5,31%	veksur	5,18%	umhugsar
5,47%	Vestara	5,31%	einans	5,17%	mat
5,46%	fremstu	5,30%	vorðið	5,17%	viðgongur
5,46%	Todi	5,30%	skránni	5,16%	tískil

Figure 9: Top 84 missing wordforms, the percentages showing the percentage of the corpus left unanalysed with a list of missing wordforms up to and including the wordform in question

Table 1: Rules and results for some CG parsers

Parser	Rules	Input	Output
North Sámi	3537	2.42	1.08
Norsk Bokmål	1964	2.13	1.17
Lule Sámi	832	2.18	1.21
Faroese	301	2.45	1.24
Greenlandic	518	2.69	1.42

unification for tags. With the set unification operator $$ it is possible to refer to a set, so that the tag that first satisfies the set must be the same as all subsequent matches of the same set. Cf. the rule (1), which refers to the set (2).

(1) SELECT $$NAGD IF (0 Det)(*1C $$NAGD BARRIER NOT-NP);

(2) SET NAGD = Nom Acc Gen Dat ;

The bulk of the rules aims at disambiguating case, number and gender within the NP. One clue as to determining the correct case is the choice of preposition, as it is for the human listener. Unfortunately, most Faroese prepositions subcategorise for more than one case. What case to choose if there is a tie is ultimately dependant upon the combination of verb and preposition. At the present stage, Fdis selects Accusative for motion verbs and change of relationship PPs, otherwise it chooses Dative.

When disambiguating running text, certain high-frequent words need special attention, both because they get multiple interpretations in the morphological component, and for their key role in the sentence. A common strategy for such words is to write specific rules just for these words. For Fdis, only approximately 15 such words have received special treatment until now, among them the pronouns *hon, vit* and the ambiguous function words *at, ið, men*. Also this is an area for improvement.

The Faroese verbal paradigm shows much homonymy. Ffst follows the practice of the reference grammars, and specifies 3 persons in the singular (also when the conjugation in question shows homonymy), but only one plural form. Naturally, disambiguating of the verbal forms rests heavily upon the person of the subject.

Mapping of grammatical functions is done on the basis of morphological cues and word order, and their disambiguation mainly on the basis of word order. The grammatical function tags are directional (the distinction @OBJ> / @<OBJ indicates whether the governing verb is to found to the right or to the left, respectively). This distinction is heavily utilised in the dependency grammar.

4 The dependency grammar

The dependency grammar quite reliably delimits NPs, and the governed constituents of P and V. Eventual errors here are due to errors in Fdis. The main obstacles for a good depencency analyses are coordination and relative clauses. Attaching appropriate constituents to the clause mother node is quite a reliable process as long as the rest of the analysis is correct. Unfortunately shortcomings in coordination and relative clause analysis, and especially the low coverage of the Ffst gives too many top nodes (2.3 alleged clausal heads per clause on average, compared to the correct 1 head/clause). Even with these shortcomings, the Fdep is already at this stage a good tool for research on basic dependency relations.

5 Evaluation

5.1 Precision and recall

The parser was tested on a small corpus of 1033 words of unseen text from a new genre (Faroese education planning). The results are shown in Table 2.

Table 2: Precision, recall, accuracy and F-ms for a test corpus

Error type	tp	fp	tn	fn
Morphology	2048	369	2501	101
Syntax	1902	515	2357	245
Dependency	724	316	0	0
	prec	rec.	acc.	F-ms.
Morphology	0.85	0.95	0.91	0.90
Syntax	0.79	0.89	0.85	0.83
Dependency	0.7	1	0.7	0.82

Thus, Fdis is work in progress

As an illustration of the Fdis output, consider Figure 10 on page 6. The two leftmost columns give the output from Ffst, with all possible readings. The third column gives the output from Fdis and Fdep, with ambiguity removed, and grammatical functions and dependency added.

5.2 Processing speed

When it comes to processing speed, it seems that the bottleneck in the system is the disambigua-

Figure 10: And the earth was waste and void; and darkness was upon the face of the deep: and the Spirit of God moved upon the face of the waters

tor. Even though it is much smaller than most CG grammars, it performs clearly worse than all the other parts of the pipeline. The reason for this might be the extensive use of set unification.

Table 3: Processing speed, measured on 100000 words of running text, on a 2,4 GHz laptop

Process	Program	Words/sec
Preprocessing	perl	10446
Morphological lookup	fst	42992
Postprocessing	perl	13017
Disambiguation	vislcg3	2042
Dependency	vislcg3	18814

6 Conclusion

The Faroese grammatical analyser presented here is still in the making. It still shows that with a modest number of CG rules, one may achive results good enough for several languaguage processing tasks. Future improvements of the analyser will concentrate upon key parts of the Ffst, upon disambiguation of complex syntactic patterns, and upon the dependency analysis of coordination and relative clauses.

References

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

Eckhard Bick. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, Aarhus.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Atro Anttila. 1995. *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*. Natural Language Processing. Mouton de Gruyter, Berlin, New York.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *13th International Conference on Computational Linguistics (COLING-90)*, pages 168–173, Helsinki.

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.

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