A constraint grammar for Faroese

Trond Trosterud
University of Tromsø
March 2, 2009

1 Introduction

The present paper presents ongoing work on an fst, a CG disambiguator and a 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.

The fst and disambiguator are available as online demos (http://giellatekno.uit.no/).

2 The Faroese analyser

The system is still not fully developed, especially shortcomings in the lexicon hampers a reliable syntactic analysis of general text genres. The Faroese morphological transducer (Ffst, based upon the lexical stock of Føroysk orðabók1) still recognises 93.5 % of all wordform tokens and 62.8 % of all wordform types in running text (the discrepancy indicates that Ffst handles common words better than rare ones). Included in the result is a name guesser. Based upon capital first letter and non-Faroese phonotax the guesser is very reliable: Of the 500 most common guesses all 500 were actually names. The results could still be better, but for certain known 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.

The disambiguator (Fdis) consists of 120 rules for morphological disambiguation, 49 mapping rules, and 46 rules for GF-disambiguation. The dependency grammar (Fdep) consists of 49 rules. With this rule set, Fdis works with an accuracy of 1.14 (an average of 1.14 analyses per word in running text), on an input where the wordforms on average had 3.01 analyses (the results were obtained on a corpus of 100000 words of newstext, previously not used for rule development, disregarding the unknown words, who had no cohorts to disambiguate).

Remarkable here is the low number of disambiguation rules. A standard CG grammar usually has several thousand rules. Partly, the low number of rules in Fdis reflects the intermediate status of the parser, and also its low accuracy (for a CG parser an accuracy of 1.14 is not a good result), but partly it illustrates the efficiency of an innovation in vislcg3, namely set 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. (??), where the case of the determiner is selected based upon any unambiguous case form within the same NP.

\[(1) \text{SELECT } \&\& \text{NAGD IF } (\text{0 Det})(\star \text{1C } \&\& \text{NAGD BARRIER NOT-NP);} \]

---

1Poulsen et al 1997. Thanks to the authors for making the lexeme stock available for me in electronic form. Without it this project would of course not have been realisable.
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 does not specify subcategorisation frames for verbs and verb + preposition combinations, this is an area for future improvement.

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.

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 dependency 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.

When it comes to processing speed, it seems that the bottleneck in the system is the disambiguator. 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 1: Processing speed, measured on 100000 words of running text, on a 2,4 GHz laptop

<table>
<thead>
<tr>
<th>Process</th>
<th>Program</th>
<th>Words/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprocessing</td>
<td>perl</td>
<td>10446</td>
</tr>
<tr>
<td>Morphological lookup</td>
<td>fst</td>
<td>42992</td>
</tr>
<tr>
<td>Postprocessing</td>
<td>perl</td>
<td>13017</td>
</tr>
<tr>
<td>Disambiguation</td>
<td>vislcg3</td>
<td>2042</td>
</tr>
<tr>
<td>Dependency</td>
<td>vislcg3</td>
<td>18814</td>
</tr>
</tbody>
</table>

3 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 achieve 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.