Tagging a morphologically complex language using heuristics

XXX XXX
Department of XXX
University of XXX
XXX
x.xxx@xxx.xx

Abstract

We describe and evaluate heuristics, a collection of algorithmic procedures, which have been developed as a part of a linguistic rule-based tagger, *IceTagger*, for POS tagging Icelandic text. The purpose of the heuristics is to mark grammatical functions and prepositional phrases, and use this information to force feature agreement where appropriate. By developing these heuristics, as opposed to the known method of developing a large set of constraint-based linguistic rules, development time of the tagging system took only 7 man months. This contradicts the widely held belief that linguistic rule-based taggers are very labour intensive. Evaluation shows that the accuracy of two of the heuristics, which guess subjects and objects of verbs, is relatively high when compared to results of parsing-based systems. Similar heuristics could be used for POS tagging texts in other morphologically complex languages.

1 Introduction

Part-of-speech (POS) tagging is the process of classifying word tokens according to their POS. Each word is assigned a string label, a tag, denoting information about the word class and morphological features. The tag is selected from a set of allowable tags, referred to as a tagset.

Tagging methods fall into two categories: data-driven methods (DDMs), (e.g. (Brill, 1995; Ratnaparkhi, 1996; Brants, 2000)) and linguistic rule-based methods (LRBMs) (e.g. (Voutilainen, 1995; XXX, 2006). In the former approach, a pre-tagged training corpus is used to obtain, automatically, information later to be used during disambiguation. In contrast, most LRBMs use hand-crafted rules for disambiguation, as opposed to information automatically deduced from corpora.

Due to the scarcity of tagged corpora for languages other than English, German and French, the usage of DDMs may not be a viable option for certain other languages. Additionally, in the case where a tagged corpus does indeed exist, data sparseness problems can occur when the size of the corpus is small, in relation to the size of the tagset (Schmid, 1995).

In our previous work, we showed that developing a LRBM for a morphologically complex language, like Icelandic, can be a feasible option (XXX, 2006). We showed that our tagger, *IceTagger*, based on this method, achieves 91.47% average tagging accuracy, when tested, using a tagset of about 660 tags, against the Icelandic Frequency Dictionary (*IFD*) corpus, a balanced corpus consisting of 590k tokens (Pind et al., 1991). This is substantially higher than the 90.36% accuracy, achieved by the best performing DDM, using the same corpus and tagset (Helgadóttir, 2004).

In this paper, we describe the heuristics used by *IceTagger*. The purpose of the heuristics is to tag grammatical functions and prepositional phrases, and use these tags to force feature agreement where appropriate. The heuristics are used by the tagger after the application of local rules, i.e. rules which perform initial disambiguation based on local context. The development of these heuristics is the main reason for a relatively short development time of the tagging system (consisting of a tokeniser, a sentence segmentiser, an unknown word guesser and a disambiguator), i.e. only 7 man months.

This paper is organised as follows. In section 2, we briefly describe the Icelandic language and its tagset. Section 3 describes different tagging methods, and, in section 4, we discuss previous work. The tagger is described in section 5, and section 6 covers the heuristics in detail. In section 7, we present an evaluation of the heuristics, and section 8 discusses the errors and refinements. We conclude, in section 9, with a summary.

2 The Icelandic language and its tagset

The Icelandic language is one of the Nordic languages which comprise the North-Germanic branch (Danish, Swedish, Norwegian, Icelandic, Faroese) of the Germanic language tree. From a syntactic point of view, Icelandic has a subject-verb-object (SVO) word order, which is, nevertheless, relatively free. The Icelandic language is a morphologically rich language, mainly due to inflectional complexity. A thorough description of the language can, for example, be found in (Þráinsson, 1994).

Due to the morphological richness of the Icelandic language, the main tagset (about 660 tags), constructed in the compilation of the *IFD* corpus, is large and makes fine distinctions. We can illustrate the preciseness of the tags by the following. Each character in the tag has a particular function. The first character denotes the word class. For each word class there is a predefined number of additional characters (at most six) which describe morphological features, like gender, number and case for nouns; degree and declension for adjectives; voice, mood and tense for verbs, etc. Table 1 shows the semantics of the noun and the adjective tags.

To illustrate, consider the sentence "*fallegu hestarnir hoppuðu*" (beautiful horses jumped). The corresponding tag for "*fallegu*" is "*lkfnvf*" denoting adjective, masculine, plural, nominative, weak declension, positive; the tag for "*hestarnir*" is "*nkfng*" denoting noun, masculine, plural, nominative with suffixed definite article, and the tag for "*hoppuðu*" is "*sfg3fþ*" denoting verb, indicative mood, active voice, 3^{rd} person, plural and past tense. Note the agreement in gender, number and case between the adjective and the noun, and the agreement in person and number between the adjective/noun and the verb.

Char #	Category/ Feature	Symbol – semantics
1	Word class	**n**–noun, **l**–adjectives
2	Gender	**k**–masc., **v**–fem., **h**–neuter, **x**–unspec.
3	Number	**e**–singular, **f**–plural,
4	Case	**n**–nom., **o**–accusative, **þ**–dative, **e**–genitive
5	Article	**g**–with suffixed article
5	Declension	**s**–strong, **v**–weak
6	Proper noun	**m**–person, **ö**–place, **s**–other proper name
6	Comparison	**f**–positive, **m**–comp., **e**–superlative

Table 1: The semantics of the noun and the adjective tags.

3 Tagging methods

Various DDMs have been developed in the last ten to fifteen years. Well known methods include probabilistic trigram methods (Brants, 2000), maximum entropy methods (Ratnaparkhi, 1996) and the transformation-based learning approach (Brill, 1995). The main advantage with the DDMs is that they are both language and tagset independent, and no (or limited) human effort is needed for rule writing. On the other hand, the disadvantage is that a pre-tagged corpus is essential for training, and a limited window size is used for disambiguation (e.g. three words in the case of a trigram tagger).

Most LRBMs use hand-crafted rules for disambiguation and are developed for tagging a specific language using a particular tagset. The advantage of LRBMs is that they do not rely on the existence of a pre-tagged corpus, and rules can be written to refer to words and tags in the entire sentence. Developing a LRBM, able to compete with data-driven taggers, has, however, been considered a difficult and time-consuming task (Brill, 1992; Samuelsson, 1994; Voutilainen, 1995).[1]

One of the better known LRBMs is the *Constraint Grammar* (CG) framework (Karlsson, 1990), in which both POS and grammatical functions are tagged. The English CG project *EngCG-2*, developed over several years, consists of 3,600 rules (Samuelsson and Voutilainen, 1997). An-

[1] A different opinion, indeed, has been expressed in (Chanod and Tapanainen, 1995).

other example of a CG project is a tagger for Norwegian which took seven man-years to develop (Hagen et al., 2000). The time effort needed in these two CG systems, for developing rules for POS tagging alone, is not available, but is probably measured in man-years. A disadvantage of the Constraint Grammar Framework is *that constraints cannot be generalised, but have to be stated in a case by case fashion* (Hinrichs and Trushkina, 2002). This is probably the reason for the large number of rules usually developed under this framework.

4 Previous work

Obtaining high tagging accuracy on Icelandic text is hard, due to the morphological complexity of the language and the large tagset used. Baseline accuracy on Icelandic text is only about 76% (XXX, 2006).

The first tagging results for Icelandic text were presented in (Helgadóttir, 2004), using the *IFD* corpus and the three data-driven taggers: *TnT* (Brants, 2000), *fnTBL* (Ngai and Florian, 2001) and MXPOST (Ratnaparkhi, 1996). The highest average accuracy, 90.36%, was obtained by the TnT tagger. By combining taggers using a simple voting scheme, i.e. selecting the tag chosen by two or more of the taggers and selecting TnT's tag in the case where all the three taggers disagreed, the total accuracy increased to 91.54%.

The 90.36% accuracy, for a single tagger, is substantially lower than has been achieved for related languages, e.g. Swedish (or English for that matter) where 93.55% accuracy was obtained in an experiment using the same taggers and a tagset consisting of 139 tags (Megyesi, 2002).

We have, previously, developed a linguistic rule-based tagger, *IceTagger*, and an unknown word guesser, *IceMorphy*, with the purpose of, first, achieving higher tagging accuracy than previously published, and, secondly, for improving the tagging accuracy using simple voting. The average tagging accuracy of *IceTagger* is 91.47%. The error rate is about 11% lower compared to using the TnT tagger. Moreover, by combining *IceTagger* with versions of *fnTBL* and *TnT*, which use features of *IceMorphy*, the tagging accuracy increased to 92.94%. The development time of the system was only 7 man months (XXX, 2006).

The error rate of the EngCG-2 system has been reported as an order-of-magnitude lower than the error rate of a statistical tagger (Samuelsson and Voutilainen, 1997). However, it is important to note that the EngCG-2 system does not perform full disambiguation and the results are, thus, only presented for the same amount of remaining ambiguity. Additionally, as previously stated, the EngCG-2 has been developed over several years.

5 The linguistic rule-based tagger

Our linguistic rule-based tagger, *IceTagger*, consists of two phases: introduction of ambiguity and disambiguation. In the former phase, the set of possible tags for each word, both known words (for which tags are sorted by descending frequency) and unknown words, is introduced. This is achieved with the help of a lexicon, automatically derived from the *IFD* corpus, and *IceMorphy*, whose function is to guess the possible tags for words not known to the lexicon. For a thorough description of *IceTagger* and *IceMorphy* the reader is referred to (XXX, 2006).

The main characteristic in the disambiguation part of *IceTagger* is the usage of only about 200 local rules along with heuristics that perform further disambiguation based on feature agreement. The purpose of a local rule is to eliminate inappropriate tags from words based on a window of 5 words; two words to the left and right of the focus word. A typical local rule uses the word class feature of surrounding tags to eliminate a particular tag of the focus word (in fact, a rule can refer to all the individual features of a tag), e.g. to eliminate a preposition tag if the following word does only have verb tags.

Henceforth, we will use the following main illustrative sentence: "*gamli maðurinn borðar kalda súpu með mjög góðri lyst*" (*old man eats cold soup with very good appetite*).[2] After introduction of ambiguity and the application of local disambiguation rules by *IceTagger*, the words of this sentence have the following tags ("_" is used as a separator between tags for a given word):

(1) *gamli*/**lkenvf** *maðurinn*/**nkeng**
borðar/**sfg3en_sfg2en**
kalda/**lhenvf_lkfosf_lveosf_lkeþvf_**
lheþvf_lheovf_lheevf
súpu/**nveo_nveþ_nvee** *með*/**aþ_aa**
mjög/**aa** *góðri*/**lveþsf** *lyst*/**nveþ_nveo_nven**[3]

[2] When translating examples to English we use word-by-word translation.

[3] **sfg3en/sfg2en**: verb, indicative, active, $3^{rd}/2^{nd}$ pers.,

6 The heuristics

Once local disambiguation has been carried out, each sentence is sent to a global heuristic module. The heuristics are used to tag grammatical functions and prepositional phrases (PPs), and force feature agreement where appropriate. We call these heuristics global because, when disambiguating a particular word, a heuristic can refer to another word which is not necessarily in the nearest neighbourhood. Each heuristic is general, in the sense that it can be applied to a sequence of words of different word classes, as opposed to the local rules which are written on a case by case basis.

Before the heuristics are applied, each sentence is partitioned into clauses using tokens like comma, semicolon and coordinating/relative conjunctions as separators (care is taken not to break enumerations up into individual parts). The heuristics then repeatedly scan each clause and perform the following: 1) mark PPs, 2) mark verbs, 3) mark subjects, 4) force subject-verb agreement, 5) mark objects, 6) force subject-object agreement, 7) force verb-object agreement, 8) force nominal agreement and 9) force PP agreement.

We will now consider each heuristic above in turn, as well as briefly describing other miscellaneous heuristics. For space reasons, we will describe the main functionality without going into too much detail or describing exceptions. Recall that, before the heuristics are run, local rules have been applied and the list of tags for each known word is sorted by descending frequency.

6.1 Marking prepositional phrases

The first heuristic searches for words, in the current clause, having a prepositional tag as their first (i.e. most frequent)remaining tag. Each such word is assumed to be a preposition and, thus, all non-prepositional POS tags for the word are removed. Additionally, the word is marked with a *PP* tag. Nominals following the assumed preposition are marked with a *PP* tag as well, if there is a case feature agreement match between the nominals and the preposition.

In (1), each word (with the exception of the adverb) in the PP "*með mjög góðri lyst*" is marked with a *PP* tag, resulting in the following POS and syntactic tags:

(2) *með*/**aþ** *PP mjög*/**aa** *góðri*/**lveþsf** *PP*
lyst/**nveþ_nveo_nven** *PP*

6.2 Marking verbs

When marking verbs in the current clause, words are searched which have a verb tag as their first remaining tag. Each such word is assumed to be a verb and, hence, all non-verb POS tags, for the word, are removed. Each verb found is marked with a functional verb tag *VERB*.

In (1), "*borðar*" is marked with the tag *VERB*.

6.3 Marking subjects of verbs

The third heuristic marks the one closest subject of a given verb, i.e. in most cases the head (a noun) of a subject noun phrase (NP). Since Icelandic word order is relatively free, both "*Jón gaf eina bók*" (*John gave one book*) and "*eina bók gaf Jón*" (*one book gave John*) are possible. The heuristic thus assumes that subjects can be found either preceding or following the verb.

For each verb *v*, already marked with a *VERB* tag, the tokens are first scanned starting from the left of *v* (since SVO order is more likely than OVS order). If the immediate token to the left of *v* is a relative conjunction or a comma, then it is assumed that the subject can be found in the previous clause (see below). Otherwise, if the current token is a nominal (not marked with a *PP* tag) and it agrees with *v* in person and number, it is marked with a functional tag *SUBJ* – if not, the scanning continues.

If no subject candidate is found to the left of *v*, a search continues using the next two tokens to the right of *v* (it is thus assumed that subjects appearing further away to the right are unlikely), using the same feature agreement criterion as before.

If at this point a subject candidate has still not been found, a search is performed in the previous clause, and the first nominal found is then marked with a functional subject tag (if it is not already marked as an object of a verb in the previous clause).

In (1), "*maðurinn*" is marked as a subject because it agrees with the verb "*borðar*" in person and number (notice that the modifier "*gamli*" is not marked – the heuristic described in section 6.8 will force an agreement between modifiers and heads of NPs), i.e.:

sing., present tense., **aþ**: preposition governing dat., **aa**: adverb. See table 1 for the semantics of the noun and the adjective tags.

(3) *gamli*/**lkenvf** *maðurinn*/**nkeng** *SUBJ*
borðar/**sfg3en_sfg2en** *VERB*

6.4 Forcing subject-verb agreement

Once verbs and subjects of verbs have been identified, feature agreement is forced between the respective words.

In (3), this means removing the second person tag from the verb "*borðar*" because the subject "*maðurinn*" is third person. Moreover, if the subject is in the nominative case (which is generally the case except for subjects of special verbs that demand oblique case subjects) all non-nominative cases are removed from the subject.

6.5 Marking objects of verbs

This heuristic marks direct objects and verb complements. Both types receive the same functional tag *OBJ*. For each verb already marked with a *VERB* tag, a search is performed for objects following the verb or, if the search is unsuccessful, for objects preceding the verb.

Objects can be nominals (direct objects or complements) or past participle verbs (only complements). When searching for nominals, words which have already been marked with *PP* or *SUBJ* tags are ignored. Only the last word in a sequence of nominals is marked. Effectively, in most cases, this means that only the head word of a NP is marked as an object. For the purpose of enforcing feature agreement between adjacent nominals, marking the head is sufficient, because, as previously stated, internal NP agreement is forced by the heuristic described in section 6.8.

In (1), the noun of the NP "*kalda súpu*" is marked as an object and the whole sentence now has the following tags:

(4) *gamli*/**lkenvf** *maðurinn*/**nkeng** *SUBJ*
borðar/**sfg3en** *VERB*
kalda/**lhenvf_lkfosf_lveosf_lkeþvf_**
lheþvf_lheovf_lheevf
súpu/**nveo_nveþ_nvee** *OBJ með*/**aþ_aa** *PP*
mjög/**aa** *góðri*/**lveþsf** *PP lyst*/**nveþ_nveo_nven**
PP

6.6 Forcing subject-object agreement

In Icelandic, feature agreement is needed between a subject and a verb complement. For example, in sentences like "*Jón er fallegur*" and "*María er falleg*" (*John/Mary is beautiful*), the complement adjusts itself to the subject. This heuristic forces such an agreement.

6.7 Forcing verb-object agreement

Icelandic verbs govern the case of their direct objects which is, generally, either accusative or dative. A verb complement is, however, always in the nominative case. The correct case of a direct object must be "learnt" for each verb, because no general rule applies. For example, "*Jón gaf bókina*" (accusative object; *John gave book*) is correct but not "*Jón henti bókina*" but rather "*Jón henti bókinni*" (dative object; *John threw book*).

A lookup table, automatically derived from the *IFD* corpus, is used for determining the correct case for direct objects (this table thus provides partial verb subcategorisation information). A lookup is performed for a given verb lexeme and the correct case is returned. Tags of the associated object that do not include the correct case are then removed. If the lookup is unsuccessful, and the marked object is not a complement, then only the nominative case tags are removed from the object (in this case, as later discussed, the most frequent tag of the object is used).

In (4), the verb "*borðar*" demands an accusative object, and, as a result, all non-accusative case tags are removed from the object "*súpu*". After this removal, the sentence part "*borðar kalda súpu*", thus, contains the following tags:

(5) *borðar*/**sfg3en** *VERB*
kalda/**lhenvf_lkfosf_lveosf_lkeþvf_**
lheþvf_lheovf_lheevf *súpu*/**nveo** *OBJ*

6.8 Forcing agreement between nominals

Agreement in gender, number and case between a noun and its modifiers is a characteristic of Icelandic NPs. This heuristic forces such an agreement in the following manner. Starting at the end of a clause, it searches for a nominal *n*, i.e. a head of a NP. If a head is found, the heuristic searches for modifiers to the left of *n* (care must be taken not to step inside a PP phrase if *n* itself is not part of that PP phrase). Agreement is forced between the head and its modifiers by removing inappropriate tags from either word.

In (5), the heuristic removes the six tags **lhenvf_lkfosf_lkeþvf_lheþvf_lheovf_lheevf** from the adjective "*kalda*", in order to force gender, number and case agreement with the tags of the following noun "*súpu*" (fem., sing., acc.). Additionally, this heuristic removes the tags **nveo_nven** from the noun "*lyst*" (see (2)) because of the feature agreement with the preceding

adjective "*góðri*". Notice that an agreement already holds in the first NP, "*gamli maðurinn*" (see (3)). After these tag eliminations, the final disambiguated sentence looks like:

(6) *gamli*/**lkenvf** *maðurinn*/**nkeng** *SUBJ*
borðar/**sfg3en** *VERB*
kalda/**lveosf** *súpu*/**nveo** *OBJ*
með/**aþ** *PP mjög*/**aa** *góðri*/**lveþsf** *PP lyst*/**nveþ** *PP*

6.9 Forcing prepositional phrase agreement

The last main heuristic forces feature agreement in prepositional phrases. Two things need to be accounted for. First, in the case when a preposition has two possible case tags, i.e. accusative and dative tags (which is common for prepositions like "*á, eftir, fyrir, í, með*" (*on, after, for, in, with*)), the heuristic removes one of the case tags based on the case of a following word in the PP.

If a following word does not unambiguously select the correct tag for the preposition then a search is performed for a preceding verb. A verb-preposition pair does, usually, unambiguously determine the correct case of the preposition. For example, in the sentence "*Jón settist á plötu*" (*John sat-down on brick*) the verb-preposition pair "*settist á*" determines an accusative case for the preposition "*á*". In contrast, in the sentence "*Jón lá á plötu*" (*John lay on brick*) the pair "*lá á*" determines a dative case for the preposition "*á*". In this case, a lookup table, automatically derived from the *IFD* corpus, is used for determining the correct case of the preposition. A lookup is performed for a given verb-preposition lexeme, the correct case returned and the conflicting tag of the preposition is removed. If the lookup is unsuccessful the most frequent tag of the preposition is used.

Secondly, once the correct preposition case tag is determined, a case agreement between the preposition and the rest of the words in the PP is forced. This is straight-forward, since the correct case is now known and the words to search for have already been marked by the heuristic described in section 6.1.

This heuristic does not have any affect on our example sentence because the sentence is, at this point, already fully disambiguated.

6.10 Other miscellaneous heuristics

In addition to the above main heuristics, specific heuristics are used to choose between supine and past participle verb forms, infinitive or active verb forms, and ensuring agreement between reflexive pronouns and their antecedents. Finally, for words that have still not yet been fully disambiguated, the default heuristic is simply to choose the most frequent tag.

Tag	Gold standard	Generated by *IceTagger*
SUBJ	265 (15.7%)	254 (15.4%)
VERB	425 (25.1%)	423 (25.6%)
OBJ	216 (12.8%)	219 (13.3%)
PP	785 (46.4%)	754 (45.7%)
Total	1,691 (100%)	1,650 (100%)

Table 2: Partition of SynFun tag types.

7 Evaluation

In this section, we evaluate the heuristics *per se*, i.e. the accuracy of the syntactic and functional (SynFunc) tagging, which the heuristics base their disambiguation process on.

We built a *gold standard* by randomly selecting 150 sentences from the *IFD* corpus and hand-tagged these sentences with SynFun tags, i.e. *PP* tags and *SUBJ*, *VERB* and *OBJ* tags. The sentences contain a total of 2,868 tokens, i.e. 19.1 tokens per sentence, on the average. During hand-tagging, 1,691 (59%) tokens received a SynFun tag.

We then ran *IceTagger* on the 150 sentences and computed precision and recall[4] for the SynFun tags generated by the tagger. The POS tagging accuracy of *IceTagger* for these sentences was 92.29%, and the ratio of unknown words was 8.26%.

Table 2 shows how the 1,691 tokens divide between the four SynFun tags. Not surprisingly, the number of *PP* tags is highest because each word in a PP (with the exception of an adverb) is tagged. Furthermore, *VERB* tags outnumber *SUBJ* and *OBJ* tags because a verb(s) occurs in almost every sentence. More *SUBJ* tags than *OBJ* tags are found which can be explained by the fact that not all verbs are transitive, but a subject is, generally, needed.

Table 3 shows precision (p), recall (r) and F-measure, $F_{\beta=1}$ (2*p*r/(p+r)), for the different tag types, guessed by the heuristics. The table shows

[4] *Precision* = # of correct generated tags / # of generated tags. *Recall* = # of correct generated tags / # of tags in the *gold standard*.

Tag	Precision	Recall	F-measure
SUBJ	85.43%	81.89%	83.62%
VERB	94.56%	94.12%	94.34%
OBJ	72.60%	73.61%	73.10%
PP	97.61%	93.76%	95.65%

Table 3: Precision, recall and F-measure for SynFun tag types, guessed by the heuristics.

much higher F-measure for *VERB* and *PP* tags compared to *SUBJ* and *OBJ* tags. This is to be expected because guessing the former is much easier than guessing the latter. As explained in section 6.2, a token receives a *VERB* functional tag if the first POS tag, in its (locally disambiguated) tag list, is a verb tag. Similarly, a preposition candidate is easy too guess and the accompanying PP words are just those nominals having the same case as the preposition. Guessing the functional *SUBJ* and *OBJ* tags is, however, more difficult because the correct guess is not only dependent on the word class, but also on word order and verb subcategorisation information.

Recall that 8.26% of the tokens, behind the figures in table 3, were unknown to the tagger. As expected, the accuracy improves by including the unknown words in the lexicon. In that case, $F_{\beta=1}$ is 85.11%, 94.95%, 74.59% and 95.73%, for *SUBJ*, *VERB*, *OBJ* and *PP*, respectively (the POS tagging accuracy is 94.25%).

8 Discussion

There are various different causes for errors in the *SUBJ* and *OBJ* tagging. One source of error is the lack of verb subcategorisation information in *IceTagger*. For example, in the sentence "*þarna svelgdist ykkur á bjórnum*" (*there quaff you on beer*) the verb "*svelgdist*" demands a dative subject (but not the usual nominative subject) and, hence, the pronoun "*ykkur*" should be tagged with a *SUBJ* tag, but not an *OBJ* tag.

Table 3 shows substantially higher F-measure for *SUBJ* vs. *OBJ*. We have noticed that PPs are responsible for many of the *OBJ* errors (and, indeed, some of the *SUBJ* errors as well). In the sentence "*hann heyrði með öðru eyranu hljóðin*" (*he heard with one ear sounds*), the noun "*hljóðin*" is a direct object of the verb "*heyrði*", but the PP "*með öðru eyranu*" lies between the verb and the object. The heuristic described in section 6.5 does not handle such intervening PPs. Furthermore, in some cases, *IceTagger* tags OBJs as VERBs, due to lack of an appropriate local disambiguation rule.

Our error analysis implies that the accuracy of *SUBJ*/*OBJ* tagging may be improved by the following. First, by adding more thorough verb subcategorisation information to *IceTagger*. Secondly, by "stepping over" intervening PPs, between the verb and the corresponding *SUBJ* or/and *OBJ*, when searching for subjects and objects. Lastly, by writing more local rules, thus eliminating more inappropriate tags before the heuristics are applied. Improving the accuracy of *SUBJ*/*OBJ* tagging will most probably increase the POS tagging accuracy of *IceTagger*.

It would be interesting to compare the figures for the functional *SUBJ* and *OBJ* tags with corresponding evaluation figures produced by a parser for Icelandic text. Unfortunately, no such figures are available.[5] Several results on tagging grammatical functions have, however, been published for related languages.

A recent study on grammatical function assignment for German (using memory-based learning from a syntactically annotated corpus with grammatical functions tags), showed $F_{\beta=1}$ as 87.23%, 78.60% and 75.32%, for subjects, accusative objects and verb complements, respectively (Kouchnir, 2004) (recall that our figures for *OBJ* tags include both direct objects and verb complements). In another German study (using finite-state cascades to annotate grammatical functions on top of a shallow constituent structure), the corresponding $F_{\beta=1}$ were 90.77%, 81.86% and 79.61%, respectively (Müller, 2004).

Since both these methods are based on parsing, higher scores are to be expected in comparison to our (non-parsing) heuristics. Nevertheless, this comparison shows that the accuracy of our heuristics for tagging subjects and objects of verbs is relatively high. Moreover, improving the accuracy of these heuristics is possible, as discussed above.

9 Conclusion

We have described heuristics used by *IceTagger*, a linguistic rule-based tagger for tagging Icelandic text. The purpose of the heuristics is to tag grammatical functions and prepositional phrases, and

[5] Indeed, only one parser for the Icelandic language currently exists. It is a parser based on HPSG, developed by a private Icelandic software company.

use these tags to force feature agreement where appropriate.

Our linguistic rule-based framework, consisting of local rules for initial disambiguation and heuristics for further disambiguation, could be applicable to other morphologically complex languages. The development of a tagging framework like ours is a feasible option when the usage of a data-driven method is difficult, due to lack of pre-tagged corpora or due to data sparseness.

By developing these heuristics, only a relatively few local constraint-based rules are needed in *IceTagger*. This is the main reason why the development time of our tagging system is only measured in 7 man months.

Acknowledgements

The author would like to thank Professor Yorick Wilks for valuable comments and suggestions in the preparation of this paper. Additionally, the Institute of Lexicography at the University of Iceland receives gratitude, for kindly providing access to the *IFD* corpus used in this research.

References

T. Brants. 2000. Tnt: A statistical part-of-speech tagger. In *Proceedings of the* 6^{th} *Conference on Applied natural language processing*, Seattle, WA, US.

E. Brill. 1992. A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the* 3^{rd} *Conference on Applied natural language processing*, Trento, Italy.

E. Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, 21(4):543–565.

J-P. Chanod and P. Tapanainen. 1995. Tagging French – comparing a statistical and a constraint-based method. In *Proceedings of the* 7^{th} *Conference on European Chapter of the ACL Conference*, Dublin, Ireland.

K. Hagen, J. Johannessen, and A. Nøklestad. 2000. A Constraint-Based Tagger for Norwegian. In C.-E. Lindberg and S. Nordahl Lund, editors, 17^{th} *Scandinavian Conference on Computational Linguistics. Odense Working Papers in Language and Communication*, volume 19, pages 31–48. Odense, Denmark.

S. Helgadóttir. 2004. Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In H. Holmboe, editor, *Nordisk Sprogteknologi 2004*. Museum Tusculanums Forlag.

E.W. Hinrichs and J.S. Trushkina. 2002. Getting a Grip on Morphological Disambiguation. In *Proceedings of KONVENS 2002, 6. Konferenz zur Verarbeitung natürlicher Sprache*, Saarbrücken, Germany.

F. Karlsson. 1990. Constraint Grammar as a Framework for Parsing Running Text. In H. Karlgren, editor, *Papers presented to the* 13^{th} *International Conference on Computational Linguistics*, Helsinki, Finland.

B. Kouchnir. 2004. Knowledge-Poor Grammatical Function Assignment for German. Manuscript. Seminar für Sprachwissenschaft.

B. Megyesi. 2002. *Data-driven Syntactic Analysis: Methods and Applications for Swedish*. Ph.D. thesis, KTH, Stockholm, Sweden.

F-H. Müller. 2004. Annotating Grammatical Functions in German Using Finite-State Cascades. In 20^{th} *International Conference on Computational Linguistics*, Geneva, Switzerland.

G. Ngai and R. Florian. 2001. Transformation-Based Learning in the Fast Lane. In *Proceedings of the* 2^{nd} *Conference of the North American Chapter of the ACL*, Pittsburgh, PA, USA.

J. Pind, F. Magnússon, and S. Briem. 1991. *The Icelandic Frequency Dictionary*. The Institute of Lexicography at the University of Iceland, Reykjavik, Iceland.

H. Þráinsson. 1994. Icelandic. In E. König and J. Auwera, editors, *The Germanic Languages*. Routledge.

A. Ratnaparkhi. 1996. A Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, Philadelphia, PA, USA.

C. Samuelsson and A Voutilainen. 1997. Comparing a linguistic and a stochastic tagger. In *Proceedings of the* 8^{th} *Conference on European Chapter of the ACL*, Madrid, Spain.

C. Samuelsson. 1994. Morphological tagging based entirely on Bayesian inference. In R. Eklund, editor, 9^{th} *Scandinavian Conference on Computational Linguistics*, Stockholm, Sweden.

H. Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *European Chapter of the ACL SIGDAT workshop*, Dublin, Ireland.

A. Voutilainen. 1995. A syntax-based part-of-speech analyzer. In *Proceedings of the* 7^{th} *Conference on European Chapter of the ACL*, Dublin, Ireland.

X. XXX. 2006. Tagging Icelandic text: A linguistic rule-based approach. Technical Report, Department of Computer Science, University of X.