Part-of-speech and Morphology Tagging Old Swedish

Gerlof Bouma, Yvonne Adesam

Dept of Swedish / Språkbanken
University of Gothenburg
gerlof.bouma@gu.se, yvonne.adesam@gu.se

1. Purpose, Background and Material

Natural language processing for historical material almost inevitably runs into the problematic combination of large variation (leading to domain adaptation-like problems) and low resources (problematic for the standard statistical methods of the field). In this paper we present our ongoing efforts in part-of-speech and morphology tagging Old Swedish, the medieval Swedish of the 13th to early 16th century. In comparison to Contemporary Swedish, Old Swedish shows syntactic differences, e.g., in its word order and argument realization, but in particular it has a richer morphology, comparable to that of contemporary German and especially Icelandic. The morphology however was in the process of becoming simpler during the Old Swedish period itself.

We have chosen to try to tackle the problem of creating a tagger for this language period using a combination of manual annotation of training material, the application of normalization strategies and the incorporation of knowledge from an external resource, namely Söderwall's (1884–1918) dictionary over the Old Swedish language.

We manually annotated a training corpus of 18k tokens (from the 14th century provincial law *Östgötalagen*) with part-of-speech tags, morphological features and lemmata. The tag set for the former two layers is described in Haugen and Øverland (2014). The latter layer consists of entries from Söderwall's dictionary. We also created three test sets of about 500 tokens each, from different periods/genres (a bit of the 13th century provincial law *Äldre Västgötalagen*, the satire *Skämtan om abbotar*, mid 15th century, and a chapter from the bible paraphrase *Pentateukparafrasen*, from a 16th century manuscript but reflecting older language).

Our work may be contrasted with that reported in Pettersson (2016), where tools for Contemporary Swedish are successfully applied to (predominantly early) Modern Swedish (16th-19th century), by adapting the application language to the tools through normalization. We expect that for our material our approach will be more fruitful, since Old Swedish differs more from Contemporary Swedish than Modern Swedish does. In addition, but related, the tag set we wish to use is different from modern tag sets, since the morphological system of Old Swedish was much richer, which makes applying tools for the contemporary language inconvenient. In the end, which is the more appropriate method is to a great extent an empirical question – its answer will have to remain future work.

	Actual		Simplified		Lemmata	
	Tok	Typ	Tok	Typ	Tok	Typ
Östgötalagen (x-val)	.11	.30	.11	.29	.05	.20
Äldre Västgötalagen	.65	.73	.50	.64	.14	.29
Skämtan om abbotar	.75	.82	.60	.76	.31	.51
Pentateukparafrasen	.79	.84	.54	.71	.35	.53

Table 1: OoV-rates for text forms (actual and simplified orthography) and lemmata, given Östgötalagen, at token and type basis.

2. Experiment setup and motivation

All experiments use the CRF-tagger Marmot (Müller et al., 2013) in its default settings: a trigram tagger without regularization, using a feature template that includes affix features for rare types.

2.1 Spelling simplication

The most obvious problem for a tagger in our material is the high proportion of out-of-vocabulary (OoV) items. As shown in Table 1, OoV-rates in the test set are as high as 80%. This is because a) the texts come from a long time span, covering different domains, which gives different (lemma) vocabularies between texts; b) the language contains an amount of inflection, which leads to many word forms per lemma; and finally c) there was no orthographic standard, so there are many spellings for the same word form.

We apply spelling simplification rules to the texts at token level to attack the last problem. The rules are intended to neutralize spelling differences that do not correspond to word form differences. Even though they both conflate too many and too few forms, they have been shown to be effective in automatic sentence segmentation for Old Swedish (Bouma and Adesam, 2013).

2.2 Lemmatization

The manual annotation includes lemma information. We experiment to which extent lemma information can overcome the OoV problem. Our manually annotated lemmata are therefore added as features in this experimental setup. The much lower out-of-lemma-vocabulary-rate is also in Table 1.

2.3 Automatic lemmatization

In a realistic scenario, manual lemmatization is not available. We therefore also investigate the use of automatically assigned lemmata. We have developed a type-based lemmatizer that links text forms to Söderwall entries using a

	Ä Västgöta		Abbotar		Pentateuk	
	Pos	Mor	Pos	Mor	Pos	Mor
No lemmata	.562	.350	.465	.301	.356	.200
With lemmata:						
Manual	.692	.442	.597	.418	.475	.281
Best from lemmatizer	.640	.401	.540	.375	.435	.264
Top 3 from lemmatizer	.723	.448	.597	.420	.495	.294
With lemmata and hints:						
Manual	.862	.576	.830	.542	.648	.380
Best from lemmatizer	.725	.483	.656	.466	.554	.335
Top 3 from lemmatizer	.756	.527	.669	.460	.535	.333

Table 2: Accuracies for POS- and morphology tagging on material in the actual spelling

	Ä Västgöta		Abbotar		Pentateuk	
	Pos	Mor	Pos	Mor	Pos	Mor
No lemmata	.707	.473	.606	.431	.537	.354
With lemmata:						
Manual	.754	.513	.677	.486	.580	.367
Best from lemmatizer	.723	.503	.608	.431	.548	.369
Top 3 from lemmatizer	.733	.511	.667	.460	.554	.356
With lemmata and hints:						
Manual	.908	.617	.826	.571	.676	.452
Best from lemmatizer	.782	.542	.712	.519	.603	.409
Top 3 from lemmatizer	.790	.554	.697	.482	.586	.401

Table 3: Accuracies for POS- and morphology tagging on material in the simplified spelling.

combination of fuzzy matching on the basis of minimal edit distance (cf Brill & Moore's, 2000, spelling correction algorithm), and a look-up list of known variants extracted from Söderwall. The edit costs for the fuzzy matching component were estimated from this same list. Fuzzy matching alone achieves a recall of .54 on held out data (.72 taking the top 3), whereas the combined method has a recall of .62 (.78 top 3) on the tagging test data.

We try two experimental setups: just adding the best guess from the lemmatizer as a feature or adding the best three guesses. For example, consider the token *cristnir* which in its context is the nominative masculine plural of the adjective **kristin**[1] 'christian'. The automatic lemmatizer's best guess for *cristnir* is however the incorrect **kristne** 'baptism', which therefore is the lemma added in the first experimental setup. The top three suggestions from the lemmatizer are **kristne** 'baptism', **kristin** 'christian' and **kristna** 'baptism/christianity' alt. 'to baptize'. Note that the latter is homonymic between a noun and a verb, but the two entries are not differentiated in the features. All three suggestions are added as features in the second setup, without indication of which was preferred by the lemmatizer.

2.4 Dictionary tagging hints

Spelling simplification and lemmatization begin to address the graphic variation problems, but they do not address the vocabulary aspect of domain variation. Söderwall's dictionary has good coverage, however, and contains information about part-of-speech and inherent morphological features, which we can exploit to improve tagging OoV-items. Söderwall's categories do not perfectly match our tag set, so we add the dictionary information as features ('tagging hints'). The tagger then learns the relation between Söderwall's categories and our target tag set.

Exactly which tagging hints are retrieved from the dictionary depends on the lemmatization strategy. For the experiments with manual lemmata, the hints are the information given for these (correct) lemmata. For instance, for **kristin**, Söderwall gives the information *adj*, which the tagger can straightforwardly associate with our part-of-speech Adjective. For the experiments with automatic lemmata, the hints are based on all entries for each of the lemmata taken from the lemmatizer. The hints are used in simple bag features: there is no information about which hint belongs to which lemma. Homonyms may lead to several conflicting hints being entered, as is the case for **kristna**, which gives *v* for the verb entry as well as *f* for the (feminine) noun.

The tagging hint *f* also shows one way in which our system deviates from Söderwall's, as we have no single label feminine noun: the tagger must learn to associate this hint with Noun in the part-of-speech layer, and Gender=FEMININE in the morphology layer. Less deterministic parts of the mapping include associating some of Söderwall's *adj* with our Quantifiers, whereas most are simply Adjectives, and to distribute Söderwall's *pron* over the different types of Pronoun and Determiner in our annotation.

3. Experimental results

The results are given in Tables 2 and 3, for actual and simplified spelling. Even for the basic setup, with no extra information added, we note that spelling simplification is very effective. Adding manual lemma information helps, as does adding the best guess from the lemmatizer, although its impact is only small for the simplified data. Adding three guesses is surprisingly effective, especially in the actual spelling experiments. Finally, adding tagging hints leads to increased performance throughout, even when retrieval is based on automatic lemmata.

Not shown in the table are cross-validation results on Östgötalagen (10 contiguous folds), yielding accuracies of .928 (part-of-speech) and .805 (morphology) in the basic setting, which shows the enormous impact of shifting domain and orthography on accuracy.

4. Future work

The combination of spelling simplification, adding lemma information and adding dictionary hints leads to an average absolute improvement of .35 tagging parts-of-speech, and .25 for morphological tagging, with manually assigned lemmata. Using a automatically assigned lemma, the improvements are still .25 and .20, respectively. The spelling simplification rules have shown to be very effective, as are

[1] In this context italics is used for mentioning word forms, (Söderwall) lemmata are set in boldface.

tagging hints from the dictionary. Lemma information itself did not add as much, although it is of course crucial for the dictionary hints.

The gap between the results with automatically assigned and manually assigned lemmata suggests that improving automatic lemmatization will also improve tagging. We plan to investigate token-based lemmatization, so that we may context information.

Looking at the errors still made by the taggers, we note two important types of errors that we have failed to consider in our experiments: proper names and numerals. The proper name tag is never used by our tagger, because their appearance is so rare in the training data. In addition, proper names are hard to recognize since they are often very specific to a document and not consistently marked by, e.g., capitalization. We will investigate the use of a name lexicon to improve this situation in the future. A second type of error concerns numerals, which may be written as Old Swedish words or as roman numerals – (parts of the) former will be in the dictionary, but not the latter. Since roman numerals are easy to recognize with good precision (some roman numerals are homographic with other words) a pre-procession round should diminish this source of error.

Acknowledgements

The research is part of the MAÞiR-project (Stiftelsen Marcus och Amalia Wallenbergs Minnesfond, MAW 2012.0146). It was previously published as Adesam and Bouma (2016).

References

Yvonne Adesam and Gerlof Bouma. 2016. Old Swedish part-of-speech tagging between variation and external knowledge. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 32–42, Berlin, Germany, August. Association for Computational Linguistics.

Gerlof Bouma and Yvonne Adesam. 2013. Experiments on sentence segmentation in Old Swedish editions. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013*, volume 18 of *NEALT Proceedings Series*.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, Hong Kong, October. Association for Computational Linguistics.

Odd Einar Haugen and Fartein Thorsen Øverland. 2014. *Guidelines for Morphological and Syntactic Annotation of Old Norwegian Texts*, volume 13(2) of *Bergen Language and Linguistic Studies (BeLLS)*.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Eva Pettersson. 2016. *Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction*. Ph.D. thesis, Uppsala University.

Knut Fredrik Söderwall. 1884–1918. *Ordbok öfver svenska medeltids-språket.* Number 54 in Samlingar utgivna av Svenska fornskriftsällskapet. Serie 1. Svenska skrifter. Svenska fornskriftsällskapet, Lund & Uppsala.