Old Swedish Part-of-Speech Tagging between Variation and External Knowledge

Yvonne Adesam
Språkbanken
Department of Swedish
University of Gothenburg
yvonne.adesam@gu.se

Gerlof Bouma
Språkbanken
Department of Swedish
University of Gothenburg
gerlof.bouma@gu.se

Abstract

We present results on part-of-speech and morphological tagging for Old Swedish (1225–1526). In a set of experiments we look at the difference between within-corpus and across-corpus accuracy, and explore ways of mitigating the effects of variation and data sparseness by adding different types of dictionary information. Combining several methods, together with a simple approach to handle spelling variation, we achieve a major boost in tagger performance on a modest test collection.

1 Introduction

Old Swedish is defined as the language stage that starts with the oldest preserved texts in the Latin alphabet (ca 1225) and ends with early print, in particular with the publication of the new testament of Gustav Vasa's bible (1526). The texts of this period are interesting as an example of a low resource and high variability material.

Compared to contemporary Swedish, Old Swedish had a different and more variable word order and a richer morphology, with nominal and verbal inflection systems resembling those of modern German or Icelandic: a nominal system with 3 genders and 4 cases and a verbal system with person and number agreement. Contemporary Swedish has 2 nominal genders,[1] at most 2 cases,[2] and no verbal agreement. Additionally, due to cultural differences and the effects of document topics/genres, the vocabulary used in Old Swedish texts may differ considerably from contemporary Swedish. We therefore expect the languages to lie too far apart to use a part-of-speech tagger trained on contemporary Swedish on Old Swedish texts.

However, until recently there have been no annotated Old Swedish texts available for training a tagger, nor any complete grammatical descriptions (i.e. computational descriptions) for inducing an annotation tool. In addition, there are a number of particularities of Old Swedish texts that are a challenge for most annotation tools and tool development methods. For example, sentence splitting cannot be handled with standard tools, as sentence boundaries are marked, if at all, in a number of ways, such as by period, slash, comma, or capitalization. Also, lack of a standardized orthography results in a wide variety of spellings for the same word, especially between texts but also within. This makes them difficult to handle with statistical methods. These problems are inflated by the fact that we are dealing with texts from wildly different genres, with different geographic origins and from a time span of roughly three centuries.

There is a long tradition of printed editions of the Old Swedish texts, for instance in the form of the editions of the medieval provincial laws by Collin and Schlyter (published 1827–1877), the publications of *Svenska fornskriftsällskapet* (The Swedish society for historical texts, 1843–present), and the *Diplomatarium Suecanum* collection of the Swedish National Archives (1820–present). More recently, electronic editions and/or electronic versions of printed editions have also become available, for instance through the Fornsvenska Textbanken project (see Delsing (2002), ∼3M tokens of Old Swedish) and the ongoing digitization efforts of the National Archives (presently ∼1M tokens of Old Swedish). The availability of such quantities of electronic text and the potential for more provides an extra motivation for our research into NLP methods for this language stage.

Although little work has previously been done

[1] Nouns only know 2 genders, adjectives may in special cases inflect for masculine in addition to common and neuter.

[2] Whether Swedish has a case distinction or not depends how one considers the genitive suffix and the subjective/objective pronominal forms.

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 32–42,
Berlin, Germany, August 11, 2016. ©2016 Association for Computational Linguistics

on automatic annotation of Old Swedish, there is related work for historical material in general. First, there exists extensive work on Modern Swedish (16th–19th c) (Pettersson, 2016, and references therein). The main difference between this work and ours is that the Modern Swedish texts are normalized to make them more similar to contemporary text, so that tools developed for contemporary material can be used. We, on the other hand, explore developing dedicated tools for the historical material by training on manually annotated historical text and using dedicated resources for the historical language variety. Since Old Swedish is more different from Contemporary Swedish than Modern Swedish is, we expect to get more mileage out of this approach than out of a transfer method.[3]

Secondly, quite a lot of work has been done for historical language variants other than Swedish, see e.g. the overview in Piotrowski (2012). Many of these also approach the historical texts by applying tools trained on the modern language variety, after adaptation of the historical texts to make them more similar to modern texts. However, for example Dipper (2011) explores normalizing the historical text to an artificial historical standard form, before training on the annotated historical text.

In this paper, we explore automatic part-of-speech (POS) tagging based on manually annotated historical text. We examine how much annotated data is needed and experiment with various ways of improving the tagging results, especially in the context of applying a tagger to documents from another domain and time. This can be achieved by handling spelling variation through a simple spelling simplification, as well as adding extra information such as manually and automatically derived lemmata, and POS and morphological information from a lexicon describing the historical language variant.

2 Materials and Tools

For our experiments, we rely on almost 20 000 tokens of text from Fornsvenska textbanken, consisting of one large and three small fragments from different texts. Around 18 000 come from the *Östgötalagen* (‘The Ostrogothic law’, based on manuscript Codex Holmiensis B50), a provincial law dating back to ∼1290 in a manuscript from ∼1350. This fragment will be used as training material. The other fragments are around 500 tokens each: the beginning of *Äldre Västgötalagen* (the ‘Elder Westrogothic law’, Cod Holm B59), the text marking the start of the Old Swedish period, dating back to ∼1220 in a manuscript from ∼1280; the complete *Skämtan om abbotar* (‘A joke on abbotts’, Cod Holm D4a), a short satire from ∼1450; and the initial chapter from *Pentateukparafrasen* (‘A paraphrase of the Books of Moses’, Cod Holm A1), from a manuscript from 1526, supposedly reflecting a text from ∼1330. These will serve as evaluation material, in part representing different genres and periods. The electronic versions of *Östgötalagen*, *Äldre Västgötalagen* and *Pentateukparafrasen* have been taken from Fornsvenska textbanken, *Skämtan om abbotar* was digitized by us from the print edition of Klemming (1887–1889).

The corpora were manually segmented, lemmatized, and annotated for POS and morphological features. We mainly followed the guidelines for Old Norwegian from the Menotec project (Haugen and Øverland, 2014), which in turn are based on the PROIEL scheme for morpho-syntactic annotation of historical text (Haug and Jøhndal, 2008). The PROIEL scheme and its associated annotation and corpus exploration environment have been used for annotating corpora of 16 other historic languages.

The manual segmentation step includes sentence segmentation, which is a non-trivial problem for automatic analysis, see Bouma and Adesam (2013), and occasionally combining or splitting graphic tokens into minimal annotation units (words). The need to combine graphic tokens into words occurs frequently for compounds which may be written as two tokens. Splitting is more rare – it is among other things needed for pronominal clitics that form one graphic token with their host. An example of a compound is *niþings værk* ‘atrocity’ in (1) below.

(1) Uerder maþer .i. kyrkiu dræpin þet ær
becomes person in church killed it is

niþings værk. þa er kyrkia al vuighz.
atrocity then is church all deconsecrated

‘If a person is killed in church, this is an atrocity, then the whole church is deconsecrated.’

We currently do not have a way of recognizing such compounds automatically. Compounds are not always clearly morphologically recognizable as such. Having an entry in one of the Old Swedish dictionaries could be taken as a pragmatic opera-

[3] A proper, direct comparison of these methods for Old Swedish will have to await future work.

tionalization of compound-hood, but because of orthographic variation, matching against a dictionary is a non-trivial matter, which we return to in the case of single-token words below. We thus use our manual segmentation as the basis in our experiments.

Example (1) also shows the use of a period in three different positions: to mark the end of a clause, the end of a sentence and to demarcate the short word *i* 'in'. Because the function and use of punctuation in the Old Swedish material varies greatly, and is not always well-understood, we remove punctuation completely for the purpose of our experiments. A similar reasoning concerns the use of uppercase, which was removed before the experiments. Finally, we also applied a light (automatic) character normalization for cases which are more at the level of character encoding than spelling differences.[4]

For the manual annotation of lemma information, we use the entries in Söderwall's (1884–1918) dictionary of Old Swedish as lemmata. New lemmata were created for those cases not covered by the dictionary, which mostly concerned names and occasionally compounds. Söderwall's dictionary is available in electronic form.[5] Lemmata, both in the form of these manually annotated gold-standard level lemmata and in the form of the output of a lemmatizer that automatically links words to entries in the electronic Söderwall, will be used in the experiments in Sections 3–5. In addition, POS- and morphology tagging hints extracted from the electronic dictionary will be used in Section 6.

We use 19 POS-tags from the PROIEL/Menotec POS-tag set and morphological features encoding person, number, tense, mood, voice, gender, case, degree, adjectival/nominal declension (definiteness). The size of the morphological tag space is about 11 500 POS-morphology combinations. In our annotated data, a total of 358 different POS-morphology combinations are used. An overview of the tagset is given in Appendix A.

For the tagging experiments we use Marmot (Müller et al., 2013), a CRF framework for large tag sets like those in morphological tagging. We use Marmot's default settings[6] and have not investigated optimization of settings and hyperparameters, instead focusing on the effects of adding/removing information on tagging accuracy.

3 Within corpus performance

We start by considering the accuracy of tagging on extremely within-domain data: data from the same corpus. This will provide us with a background to interpret the cross-document (both within and outside-of domain) results. All results will be reported for both full morphological tagging (assigning both POS-tag and morphological features) and the less fine-grained task of POS-tagging. In this paper, all averages are arithmetic means and macro averages.

3.1 Cross-validation

Cross-validation results of training and evaluating a basic model on Östgötalagen, with only the token layer as information, are given in Table 1. The table gives averages over different cross-validation regimes to get an idea of the homogeneity within the corpus as seen from the tagger. When randomly spreading sentences over ten data splits (10-fold random), the model will have seen material from all parts of the corpus, and if there are any differences with Östgötalagen that affect tagging, like systematic changes in orthography or vocabulary, these will be evened out in this way of evaluating. The tagger reaches an average POS-tagging accuracy of 94.2% under this regime, with relatively minor differences between the folds.

By taking ten consecutive parts from the corpus as splits, we get the '10-fold contiguous' regime. There is now a possibility that the tagger is confronted with evaluation data sections of the corpus it hasn't seen before. Performance drops a little bit, to 92.8%. We interpret this as an indication that the tagger has relatively little trouble generalizing to different parts of the corpus, a sign that the corpus is rather homogeneous. Note that the differences between folds has increased, with the minimum belonging to the fold with test data from the beginning of Östgötalagen.

Finally, we try to maximize the differences between folds by defining them on the text structure. Each split now corresponds to one of the

[4] In particular we neutralized the differences between *æ* and *ä*, *ø* and *ö* and *þ* and *ð*. Note that usage of *ð* is very rare in Old Swedish material, and *þ* may encode voiced as well as unvoiced dental fricatives.

[5] `https://spraakbanken.gu.se/resources`

[6] In the default settings, Marmot trains a trigram model without any regularization. Morphological tags are split into their parts by the tagger rather than treating them as atomic. The tagger automatically creates suffix and prefix features based on the token input layer. It will not predict morphological labels not seen in the training data.

		Min	Mean	Max
POS	10-fold random	.931	.942	.947
	10-fold contiguous	.897	.928	.958
	4-fold per chapter	.893	.915	.924
Morph	10-fold random	.819	.832	.841
	10-fold contiguous	.725	.805	.864
	4-fold per chapter	.751	.787	.808

Table 1: Cross-validation results for the basic model on Östgötalagen under different regimes.

major subdivisions of the legal text, the so called *balk*. We only use the four largest from our annotated material, each 3,500 to 5 000 tokens, to avoid large variations in training data size between folds. The average performance drops further to 91.5%. The lowest accuracy is achieved on the first balk, *Kyrkobalken*, concerning the church – in agreement with the 10-fold contiguous regime. We are not aware of any obvious differences, like provenance, that might explain this.

The picture for morphological tagging is the same as for POS-tagging, with average accuracy between 11 and 13 percentage-points lower. The drop in accuracy between regimes is a bit larger than for POS-tagging, meaning that the tagger is more sensitive to corpus differences in this task. It seems likely that this is directly related to the larger tag set and therefore increased data sparseness.

3.2 Lemmata and spelling

A major obstacle when working with historical text is spelling variation. For Swedish, there was no written standard until several hundred years after the Old Swedish period. When training a parser or any other statistical natural language processing tool, spelling variation leads to data sparseness, which for instance presents itself in the form of very high out-of-vocabulary (OOV) rates and large amounts of features that have to be weighted on the basis of low counts.

In this paper we investigate two orthogonal ways of remedying this: First, we add a word's lemma as a feature. We might expect this to have more of an effect on POS-tagging than on morphology tagging, as the lemma in it self does not provide explicit information about the morphology in the way for instance inflection does.[7] In this and the next section, we use the manually annotated lemmata as features, in Section 5 we investigate the effectiveness of adding the output of an automatic lemmatizer.

Secondly, we apply the spelling simplification method described in Bouma and Adesam (2013), which uses a handful of rewriting rules intended to remove differences between spellings. For instance, it replaces many repeated characters by a single character (e.g. *aa* → *a*), removes a restricted number of digraphs (e.g. *ck* → *k*, *gh* → *g*) and reduces certain characters denoting similar sounds to one (e.g. *u*, *v*, *w* → *v*). We have previously shown this crude method to be effective in a sentence segmentation task (ibid), even though the simplification can easily conflate words that are not spelling variants and at the same time may fail to bring obvious variants together.[8] The simplification rules are directly applied to the token layer. Unlike adding a lemma, spelling simplification thus strictly removes information.

Simplifying the spelling has no effect on the within-corpus tagging accuracy on Östgötalagen (average for POS remains at 92.8% under the 10-fold contiguous regime, morphology is at 80.4%), adding a lemma gives a nominal increase (93.9% POS, 81.2 morphology). Combining the two does not lead to a change with respect to just adding a lemma (see Figure 2 in Section 4).

The absence of any real effect does not come as a surprise, given our remarks about the homogeneity of the corpus in the section above. In addition, it is interesting to note that the spelling simplification has only little impact on the lexical statistics of Östgötalagen: the average OOV-rate in cross-validation is basically unaffected (see also Table 2 in Section 4), and the number of rare types, with a token frequency of ≤ 10, drops with only 2.5% points, even though 55% of the types are affected by the spelling simplification.

3.3 Training data size

Figure 1 shows the effect of training size on accuracy, for using the train-test split of one of the 10-fold contiguous folds. Within-corpus learning curves are interesting from the point of view of

[7] One of the design goals of the Menotec scheme is to avoid having the same lemma with different POS-tags, which makes having the lemma particularly informative for a POS-tagger.

[8] Spelling simplification therefore shares characteristics with stemming: a fast method to reduce variation in a corpus. But whereas stemming mainly reduces variation due to inflection and derivation, spelling simplification maintains morphological information and aims at reducing variation due to orthographic variation.

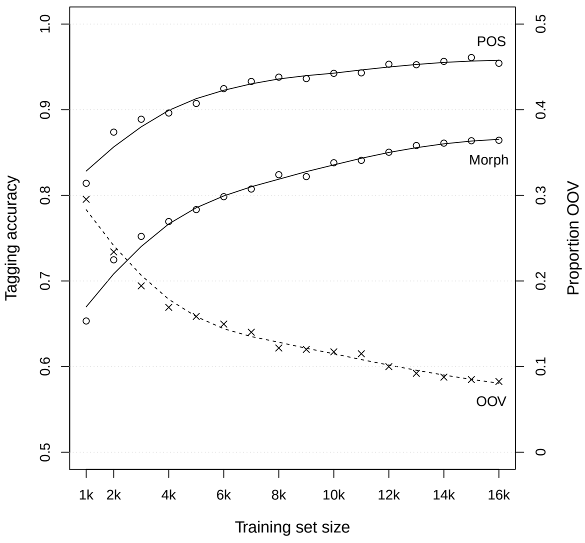

Figure 1: Learning curves for within-corpus accuracy (left axis) and related OOV proportions (right axis) on Östgötalagen for the basic model.

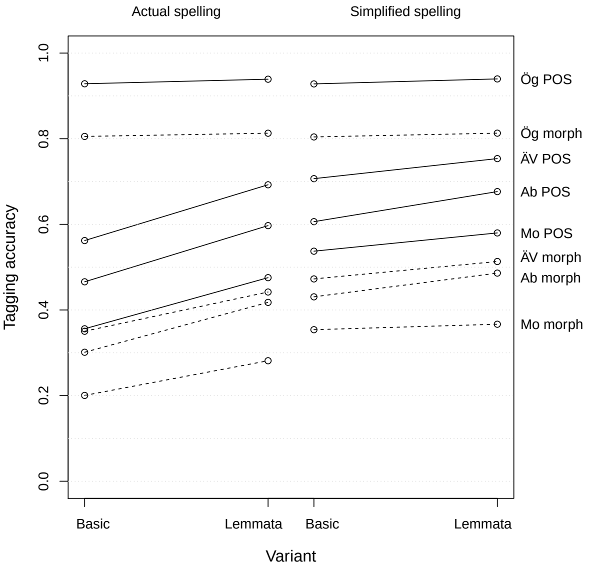

Figure 2: Effect of spelling simplification (left vs right column) and adding lemma information on tagging accuracy.

tagger-assisted manual annotation. Earlier studies on this topic (see Fort and Sagot (2010) on Penn Treebank-style POS-tagging, and Skjærholt (2011) on tagging Latin morphology using the PROIEL tagset) show that a pre-tagging accuracy of .8 and upwards can be beneficial to manual annotation speed and (to a lesser extent) accuracy, although the effect is stronger for less experienced annotators. For our Östgöta corpus, it would thus seem that a tagger trained on as little as 1 000 tokens (around one week's work for a medium-experienced annotator annotating POS, morphology and lemmata) can be of help for POS-tagging, and 7 000 tokens for morphology tagging.

4 Lemmata and spelling simplification across corpora

Let us now turn to the cross-document experiments, which will give us a better picture of what happens when we automatically annotate new texts. We train on the whole Östgötalagen data and evaluate the models on the three other texts, Äldre Västgötalagen (ÄV), Abota (Ab), and Moses (Mo) from Pentateukparafrasen. The results can be compared to the average results when performing tenfold evaluation on Östgötalagen (Ög).

The results are in Figure 2 (see Appendix B for the actual numbers). As we can see, while spelling simplification and lemma information does not help much when tagging Östgötalagen (as stated in Section 3.2), we get a large improvement from both approaches when tagging other texts. Interestingly, in all cases, spelling simplification on its own contributes more than providing the tagger with the correct lemma, for both POS-tagging and morphological tagging. Combining spelling simplification and providing a lemma gives the best results, suggesting the enhancements supply complementary information.

Over all, we get a large increase in accuracy, rendering a quite acceptable POS-tagging accuracy for all texts, if we consider a semi-automatic annotation process where we automatically tag Old Swedish text before manually checking it. Morphological tagging is lagging behind, as is to be expected, as it is a more difficult task because of the larger tag set. However, a particular problem is the occurrence of unseen morphological labels in the testing data, which because of the used tagger settings cannot be predicted correctly. For Äldre Västgötalagen, Abota and Moses, the proportion of types with an unseen morphological label is 7%, 4% and 15%, respectively.

Let us also look at the improved number of tokens (i.e. the change in number of correct tokens) between the basic tagging, without extra information, and tagging with both lemma and spelling simplification, per POS. For Äldre Västgötalagen we have a larger change (more than 10 tokens, i.e., more than 2% of all tokens) for conjunctions,

	Actual		Simplified		Lemmata	
	Tok	Typ	Tok	Typ	Tok	Typ
Ög	.11	.30	.11	.29	.05	.20
ÄV	.65	.73	.50	.64	.14	.29
Ab	.75	.82	.60	.76	.31	.51
Mo	.79	.84	.54	.71	.35	.53

Table 2: OOV-rates for words (actual and simplified orthography) and lemmata, given Östgötalagen.

nouns, and verbs, while Abota has a large change for conjunctions, adverbs, nouns, and prepositions. For Moses we see a large change for conjunctions, nouns, demonstrative pronouns, and prepositions.

For conjunctions and prepositions, most improvements come from the spelling simplification, while nouns get their improvement from both lemma and spelling simplification. Verbs also get their improvement from both, but to a larger extent from lemma. The improvements for demonstrative pronouns come from the lemma. These results are not surprising. While we get an overall large improvement from spelling simplification, lemma may be more helpful for inflected POS categories.

Exploring the data further, one reason for the difference in impact of spelling simplification and lemma may be the rate of out-of-vocabulary words (OOV) between the texts. The OOV-rates for the different test sets are given in Table 2. Not surprising, the rate of OOV is lowest for Östgötalagen, since the test data comes from the same text as the training data. The OOV-rate in Äldre Västgötalagen, being the closest to the training data in genre, is a lot higher. Abota and Moses have the highest levels of OOV for the actual spelling. However, while the spelling simplification significantly lowers the OOV-rates for all texts but Östgötalagen, it has the largest impact on the OOV-rates for Moses, lowering the percentage of OOV by 25 percentage-points at token-level and almost 15 percentage-points at type-level.

5 Automatically assigned lemmata

We have seen that adding lemma information has a beneficial effect on tagging accuracy across corpora. In a realistic setup, we do not have access to gold standard lemmata. This raises the question whether automatically assigned lemmata also will boost accuracy. To this end we have implemented a simple lexicon linking method, which assigns one or more lemmata from Söderwall's dictionary to each token. Before discussing the effect of using automatically assigned lemmata, we describe our lexicon linking strategy.

5.1 Linking tokens to lemmata

Many entries in Söderwall's dictionary contain a list of form variants, to illustrate – rather than fully document (Djärv, 2009) – the different forms due to inflection and orthographic convention. In our electronic version, we have a total of 24 000 form variants for 8 000 (out of 27 000) lemmata. A straightforward linking strategy uses these as a simple look-up table. A token is linked to any lemma that a) matches the token exactly, or b) lists a form variant that matches the token exactly. We rank multiple lemmata in this order and use alphabetical order as a further tie breaker.

Average linking scores (i.e. recall) of this method on our four corpora is given in Table 3. We see that considering only the best suggestion from the dictionary retrieves a correct lemma for 45% of the tokens (28% of types). Considering whether the correct lemma is among all returned matches raises the score, but it remains low. The reason for this is the low proportion of cases in which this method applies, that is, the cases when we get a link to the dictionary at all (61% tokens, 42% types). This low application rate motivates a combination with a method with higher recall, like a fuzzy matching-based approach that assigns a lemma to every token. Pettersson (2016) and Bollmann (2013) have shown the effectiveness of a combination of look-up and fuzzy matching for different historical languages.

Our fuzzy matching method builds on Adesam et al. (2012). A word form is matched against the lemma that gives the lowest weighted edit distance, where edit operations may map several characters at once. Edit costs are calculated from the form variants listed in Söderwall's dictionary as follows: First, each variant is character aligned with its lemma using the EM specification given in Oncina and Sebban (2006).[9] In a second step, sequences of character mappings are taken from these alignments to give counts of n-to-m-gram mappings. Source and target sequences do not have to have the same effective length, as either of them may contain ϵ-s. Finally, we assign a cost

[9] For convenience, we use a hard-EM variant of Oncina and Sebban's method. See also Wieling et al. (2012) for a similar iterative method to obtain character alignments.

		Tokens	Types
Dictionary look-up	best	.45	.28
	all	.54	.33
	applies	.61	.42
Edit distance	best	.54	.48
	top 3	.69	.67
Combo	best	.62	.55
	top 3	.78	.73
Coverage		.92	.91

Table 3: Lexicon linking scores per method and dictionary statistics

of $-\log p(\text{target}|\text{source})$ to each mapping. For our final model, we include edits that map up to 5 characters. On a held-out development set from the dictionary listed form variants, this method retrieves the correct lemma 54% of the time, with the correct lemma being among the best 3 in 72% of the cases. Models that allowed wider edits did not give clear improvements on the held-out data.

As shown in Table 3, the model retrieves the correct lemma for 54% of the tokens (48% of types) in our corpora when considering the best match only. Among the top 3 of matches, the correct lemma is found 69% of the time (67% at type level). The Moses text is an outlier here with a mere 47% token score (43% types; neither shown in the table) for the best match. Its low linking accuracy must be explained from the high incidence of proper names (see also Section 6). Indeed, this is also reflected in the low coverage of our lemma list with respect to the text, which is up to 22 percentage-points lower than for the other texts (token- and type-level).

We combine these two methods by first taking all lemmata from the dictionary look-up method, and then adding the ranked lemmata from the edit distance method. This combined approach finds the correct lemma for 62% of the tokens (55% of types). The correct lemma is among the 3 best candidates in 78% of the cases (73% type-level).

5.2 Tagging with automatically assigned lemmata

We automatically add lemma information using the method just described as features in the test and training data. We explore two ways of adding lemma information: using only the single top-ranked lemma, and taking the top 3 suggestions so that each token receives multiple possible lemmata. In the latter case, the three suggestions are values of the same key, so that the model cannot distinguish for a given lemma whether it is the first or third ranked suggestion.[10]

As before, we compare tagging results using the data in its actual spelling and in a simplified version. The results are summarized in Figure 3 (see also Appendix B). Compared to a model with access to manually annotated lemma information, a model with a single automatic lemma loses tagging accuracy, both on the POS and the morphology tasks. This effect can be seen both in the actual spelling and the simplified spelling versions, although the effect is smaller for the Äldre Västgöta and Moses subcorpora in the simplified spelling experiment.

Interestingly enough, in the actual spelling version, using multiple automatically assigned lemmata not only improves upon using a single automatically assigned lemma but also upon using the manually assigned lemma. We do not currently understand the nature of this effect, especially since it disappears in the simplified spelling setup. We hope future investigations will give us better insight into this matter.

Overall, adding automatically assigned lemmata does not hurt performance (0–5 percentage points improvement for simplified spelling) and may potentially be very helpful (5–10 percentage points for actual spelling). Most importantly, however, having a lemma gives us access to more detailed and useful information from the dictionary, as we will see in the following section.

6 Adding tagging clues from the lexicon

Entries in Söderwall's dictionary contain information about POS and in some cases information pertaining to morphological properties. This information may be the POS itself (e.g. *adv.* for an adverbial or *v.* for verb), but it may also just give us e.g. the gender for a noun (*m.* for masculine). In some cases we get further specifications (e.g. *pron. pers.* for a personal pronoun or *adj. komp.* for an adjective in comparative form). Although Söderwall's label inventory is not directly mappable to ours, we can use this information as tagging clues by including Söderwall's labels as features in the data (cf Müller et al., 2013).

[10] We also experimented using different feature keys for the first, second and third suggestion. This gave similar but slightly worse results.

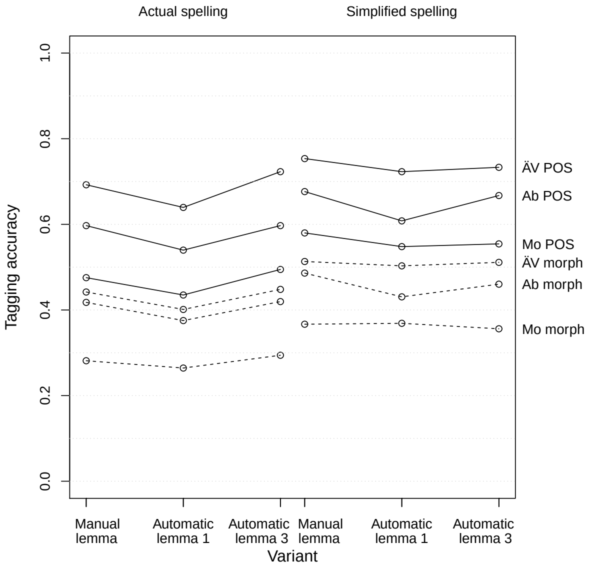

Figure 3: Tagging accuracy with manually versus automatically added lemmata.

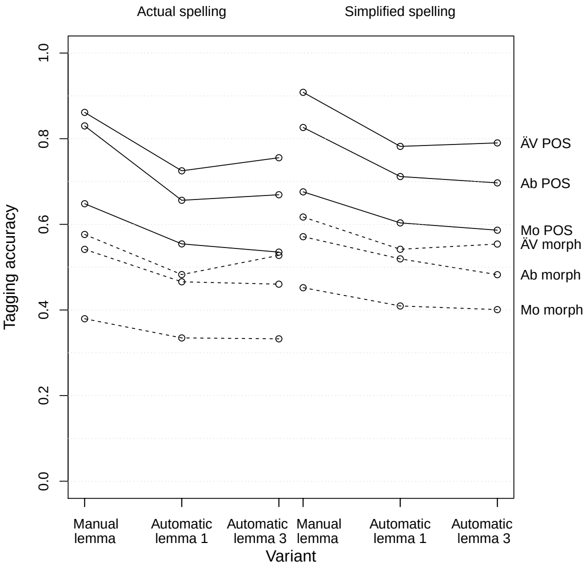

Figure 4: Tagging accuracy for manually versus automatically added lemmata with tagging hints.

We derived this information on the basis of the lemmata for each of the previous setups (manually assigned lemma, single automatically assigned lemma, multiple automatically assigned lemmata). When we have multiple lemmata for a token, we may get multiple tagging clues from the dictionary. Söderwall's dictionary may also give multiple labels, e.g. for homonyms. We include all possibilities as features with the same key.

We can extract at least one tagging clue per token (manually assigned lemma) in most cases, except for the Moses text, where we only have a coverage of ∼75%, due to proper names and numerals.

The results of using these extra tagging clues in POS and morphology tagging can be found in Figure 4 (see also Appendix B). Overall, accuracy goes up compared to the models without tagging clues (see Figure 3). Here it is clear that the models with manually assigned lemmata fare much better than those with automatically assigned lemmata. The previously seen advantage of having multiple automatically assigned lemmata has disappeared. As in each of the previous experiments, using simplified spelling improves accuracy.

On average, the best model without any manual input in the test data achieves 69.9% accuracy on the POS task and 49.0% on the morphology task (single automatic lemma with tagging clues, spelling simplification). This is a huge improvement over the initial 46.1% POS and 28.4% morphology (no lemma, actual spelling).

As mentioned, Moses achieves lower scores because it contains a lot of proper names: 72 occurrences (15% of the tokens) compared to one or two in the other two shorter texts, and 12 in the much longer Östgötalagen. Not only are individual proper names OOV, but the tagger assigns a very low probability to the word class as a whole. Indeed, the tagger never predicts the proper name label for any token in the evaluation, even under the best model. For Moses, this means that 15% of the tokens cannot be correctly tagged. Correcting only names would boost accuracy to almost 83% for POS-tags, on par with results for the Abota text.

Two other clearly problematic POS-tags are demonstrative pronouns and quantifiers. Demonstrative pronouns are tagged with low precision and recall in the Abota and Moses texts, in particular when using automatically derived lemmata. The quality of the automatically assigned lemmata cannot be the sole explanation for this effect, as it is fairly good for Abota, whereas it is low for Moses.

The label of quantifier is not only used for items expressing meanings like *all*, *each* and *some*, but also for cardinal numerals. In Moses, most of the numbers are written using roman numerals, which our tagger currently does not recognize. In Abota, it is the low quality of the automatic lemma assignment that causes problems specifically for this category. A possible reason for this is the irregular inflection paradigms for these items.

7 Conclusions

In this paper we have explored several approaches to automatic annotation of POS and morphology for Old Swedish text. These approaches have mainly been linguistically informed, and we have shown that adding clues about lemma and morphological information from a dictionary greatly improves results, together with a simplistic method for removing spelling variation.

With a training set of less than 18 000 words, we start out with an average accuracy of around .45 for coarse POS-tags (less than .30 with morphological classification) when testing on other texts. The overall best final results give us an average of .80 for POS-tags (.55 with morphological classification), using spelling simplification, manually annotated lemmata, and morphological information from the dictionary based on those lemmata. The best results with automatically induced extra information were .70 for POS-tags (.50 with morphological classification), when a single lemma was automatically selected, together with spelling simplification and morphological information from the dictionary based on the automatically extracted lemma.

We have also seen that a fairly small amount of manually annotated data, maybe as little as 1 000 words, is necessary for training a POS-tagger for aiding manual annotation, although more, above 7 000 words, is necessary for a morphology tagger.

Comparing results between the within-corpus and across-corpus experiments, we find it striking that even at the smallest within-text data set size (1 000 tokens), accuracy lies well above the accuracy of the basic model in the across-corpus setup. It is even slightly better than our best model using automatically assigned lemma information on Äldre Västgötalagen. The within-corpus learning curve underlines the severity of the differences between corpora.

We have seen that, on the one hand, spelling simplification gives better tagging results across corpora than adding lemmata, while on the other hand lemma OOV-rates are much lower than simplified spelling word OOV rates. The rate of OOV is therefore clearly not the only reason for low tagger performance across corpora. An important difference lies in the ways we added the lemmata and simplified spelling. The former was added as a feature linked to a single token, whereas for the latter we changed the token layer itself. This means that the simplified spelling also affected the suffix-/prefix-based features and the token context features the CRF tagger constructs automatically under the default settings we used. It seems plausible that this difference makes the simplified spelling much more effective. More experimentation is needed to see if lemma information is more effective when derived features are also added to the model. In any case, the effectiveness of simplified spelling also suggests that investigating proper spelling normalization may be well worth the effort.

Acknowledgments

This research is funded by Stiftelsen Marcus och Amalia Wallenbergs Minnesfond (project "MAÞiR", nr MAW 2012.0146).

References

Yvonne Adesam, Malin Ahlberg, and Gerlof Bouma. 2012. *bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa...* Towards lexical link-up for a corpus of Old Swedish. In Jancsary, editor, *Empirical Methods in Natural Language Processing: Proceedings of KONVENS 2012 (LThist 2012 workshop)*, page 365–369, Vienna.

Marcel Bollmann. 2013. Spelling normalization of historical German with sparse training data. Technical report, BLA: Bochumer Linguistische Arbeitsberichte 13.

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*.

Lars-Olof Delsing. 2002. Fornsvenska textbanken. In Lagman, Olsson, and Voodla, editors, *Nordistica Tartuensia 7*, pages 149–156, Tallinn. Pangloss.

Stefanie Dipper. 2011. Morphological and part-of-speech tagging of historical language data: A comparison. *Journal for Language Technology and Computational Linguistics, Special Issue: Proceedings of the TLT-Workshop on Annotation of Corpora for Research in the Humanities*, 26(2):25–37. `http://www.jlcl.org/2011_Heft2/2.pdf`.

Ulrika Djärv. 2009. *Fornsvenskans lexikala kodifiering i Söderwalls medeltidsordbok [The lexical codification of Old Swedish in Söderwall's medieval dictionary]*. Number 91 in Samlingar utgivna av Svenska fornskriftsällskapet. Serie 1. Svenska skrifter. Svenska fornskriftsällskapet, Uppsala.

Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development.

In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 56–63, Uppsala, Sweden, July. Association for Computational Linguistics.

Dag Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the Old Indo-European bible translations. In Caroline Sporleder and Kiril Ribarov, editors, *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, pages 27–34.

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*.

Jose Oncina and Marc Sebban. 2006. Learning stochastic edit distance: Application in handwritten character recognition. *Pattern Recognition*, 39(9):1575–1587.

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

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Morgan & Claypool.

Arne Skjærholt. 2011. More, faster: Accelerated corpus annotation with statistical taggers. *JLCL*, 26(2):153–165.

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.

Martijn Wieling, Eliza Margaretha, and John Nerbonne. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2):307–314.

A Overview of the Menotec POS-tagset

Part-of-speech	Morph features
Noun	gender, number, case, definiteness
Proper noun	gender, number, case, definiteness
Adjective	degree, gender, number, case, definiteness
Personal pronoun	case
Reflexive pronoun	case
Interrogative pronoun	gender, number, case
Indefinite pronoun	gender, number, case
Demonstrative pronoun	gender, number, case
Quantifier	gender, number, case
Possessive pronoun	gender, number, case
Verb	finiteness, tense, mood, person, number, voice
Adverb	degree
Interrogative adverb	–
Preposition	–
Coordinator	–
Subordinator	–
Interjektion	–
Unanalyzed	–
Foreign word	–

Based on Haugen and Øverland (2014).

B Overview of experimental results

	ÄV		Ab		Mo	
	Pos	Mor	Pos	Mor	Pos	Mor
Basic	.562	.350	.465	.301	.356	.200
With lemmata:						
Manual	.692	.442	.597	.418	.475	.281
Auto 1	.640	.401	.540	.375	.435	.264
Auto 3	.723	.448	.597	.420	.495	.294
With lemmata and hints:						
Manual	.862	.576	.830	.542	.648	.380
Auto 1	.725	.483	.656	.466	.554	.335
Auto 3	.756	.527	.669	.460	.535	.333

Accuracies for POS- and morphology tagging on material in the actual spelling.

	ÄV		Ab		Mo	
	Pos	Mor	Pos	Mor	Pos	Mor
Basic	.707	.473	.606	.431	.537	.354
With lemmata						
Manual	.754	.513	.677	.486	.580	.367
Auto 1	.723	.503	.608	.431	.548	.369
Auto 3	.733	.511	.667	.460	.554	.356
With lemmata and hints						
Manual	.908	.617	.826	.571	.676	.452
Auto 1	.782	.542	.712	.519	.603	.409
Auto 3	.790	.554	.697	.482	.586	.401

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

ÄV: *Äldre Västgötalagen* (490 tokens)
Ab: *Skämtan om abbotar* (541 tokens)
Mo: *Pentateukparafrasen* (469 tokens)