bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa...
Towards lexical link-up for a corpus of Old Swedish

Yvonne Adesam **Malin Ahlberg** **Gerlof Bouma**
Språkbanken
Department of Swedish
University of Gothenburg
firstname.lastname@gu.se

Abstract

We present our ongoing work on handling spelling variations in Old Swedish texts, which lack a standardized orthography. Words in the texts are matched to lexica by edit distance. We compare manually compiled substitution rules with rules automatically derived from spelling variants in a lexicon. A pilot evaluation showed that the second approach gives more correct matches, but also more false positives. We discuss several possible improvements. The work presented is a step towards natural language processing of Old Swedish texts.

1 Introduction

Corpus linguistics makes it possible to explore many interesting questions by investigating large amounts of text. These texts generally have some form of annotation, or mark-up. Språkbanken,[1] the Swedish language bank, has a large collection of modern texts with close to a billion words, which have been automatically annotated and are searchable through the corpus query interface Korp (Borin et al., 2012).[2] A current effort to cover different types of Swedish (Borin et al., 2010) involves older variants of Swedish texts, from the 19th century back to the 13th century. We now focus on Old Swedish, which covers two stages of Swedish: Early Old Swedish (1225-1375), and Late Old Swedish (1375-1526). Ultimately, our goal is to develop tools for various levels of annotation, such as part-of-speech, morphosyntactic information, and dependency parses.

A number of issues are problematic for annotation of Old Swedish texts. For example, sentence splitting cannot be handled with standard tools, as sentence boundaries are often not marked by punctuation or uppercase letters. Compared to modern Swedish texts, the Old Swedish texts have a different vocabulary and richer morphology, show a divergent word order, and Latin and German influences. Finally, the lack of a standardized orthography results in a wide variety of spellings for the same word. In the following, we will discuss our first efforts to handle such spelling variations.

2 Resources

Our source material comes from Fornsvenska textbanken,[3] a collection of around 160 digitized texts totaling 3 million words, mainly from the 13th to the 16th century. The data set consists of novels, poems, laws, and religious texts, ranging from 200 to 200,000 words in length. The texts thus vary in age, language, size, and type.

Språkbanken develops and makes available a range of lexical resources (Borin et al., 2010). The central resource for Contemporary Swedish is SALDO (Borin et al., 2008). Dalin's dictionary (1853/1855) covers 19th century Swedish. Additionally, we have three dictionaries for Old Swedish: Söderwall and Söderwall's supplement (1884/1953) with 44,000 entries, and Schlyter (1887) with 10,000 entries, focusing on law texts. There is overlap between these three dictionaries. Finally, a morphology for modern Swedish has been developed at Språkbanken and morphologies for 19th century Swedish and Old Swedish are under development (Borin and Forsberg, 2008). While the work presented only uses the Old Swedish dictionaries and corpora, we intend to also use the morphology in the future.

[1] http://spraakbanken.gu.se
[2] http://spraakbanken.gu.se/korp/
[3] http://project2.sol.lu.se/fornsvenska/

3 Spelling Variation

A problem that severely hampers processing of historical text is the lack of a standardized orthography. A common solution is to normalize spelling to a modern language variety (Pettersson et al., 2012; Bollmann et al., 2011; Jurish, 2010a). In contrast, we intend to retain the original orthography in order to use the lexical and morphological resources available. Thus, we investigate mapping forms occurring in the corpora to the entries in the Söderwall and Schlyter dictionaries. This linking is motivated by the following: the dictionary entries can serve as type identifiers, which facilitate statistical processing; we gain access to the (partial) information about part-of-speech provided in the lexica; and finally, from a user-interface perspective, the link can serve as an assisting technology for studies of the Old Swedish texts.

Initial experiments suggest that only roughly one quarter of the types in unprocessed text match exactly against a lexical entry in one of the dictionaries. We therefore investigate approximate string matching to raise the coverage of the dictionaries. In the following, we give a general outline of our method and report on a pilot evaluation. As this is work in progress, we mainly focus on the qualitative findings of this evaluation and discuss directions for future work.

3.1 General Approach

A lexical entry is considered to be a match for a token in the text if it is the best candidate in terms of edit distance below a given threshold. We only allow the substitution operation, as this simplifies the implementation and the automatic collection of rules. Substitution rules can contain strings of unequal length.

Calculating the edit distance between a source word and each of the about 54,000 target dictionary entries is too expensive with million word corpora. We therefore use an *anagram hashing* filter (Reynaert, 2011) to cut down the number of exact edit distance calculations. The filter computes character based hashes for the source words and for the target entries. Character edits such as substitution can be performed numerically on these hashes. After applying a numerical edit on a source hash, we have a new hash value which we can compare against the hash table representing the dictionary or apply further edits to. The hash function loses information about the order of characters in a string – hence the name *anagram hash* – which means that the filter underestimates the actual number of required edits between source and target is. In the experiments presented below, the filter generates, for each source word, up to 1 million hash variations with an increasing number of edits. From the target dictionary entries found in this candidate set – a much smaller number – we select the best matches using exact edit distance calculation. In its current implementation, the filter may discard valid matches, a problem we shall return to in Section 3.4.

3.2 Substitution Rules

We created sets of substitution rules in two ways. First, we composed a set of 31 correspondences by hand, based on common alternations appearing in the Old Swedish texts. An illustration of these rules is given in Fig. 1 (left). The rules have a constant weight, which makes them more costly than identity mapping but cheaper than a substitution not defined in the rule set.

Secondly, we used character aligned spelling variants to derive two further substitution rule sets. The variants are taken from the Schlyter dictionary, where 1,478 entries have an average of 2.2 variants. Note that the dictionaries do not share all orthographic conventions, and we cannot assume that all possible spellings are listed for an entry.

The spelling variants were aligned with iterated Levenshtein distance alignment (Wieling et al., 2009), using weighted substitution, insertion or deletion operations. The operations initially have unit cost, but after a complete iteration we use the alignments to calculate new weights, which are then used to realign the variant pairs. The weights converge in 3 or 4 iterations, resulting in fewer alignments, which are of higher quality. For instance, the pair *alita-ålijthe* 'be assured' can be aligned in the following two ways if all operations have unit cost:

1	a	l	i		t		a	
	å	l	i	j	t	h	e	
2	a	l	i		t	a		
	å	l	i	j	t	h	e	

After convergence of the alignment weights, however, only the former alignment is produced, as

Rule	Cost	Rule	Cost	Rule	Cost
v↔u	0.1	o↔u	0.14	u→o	0.20
v↔w	0.1	d↔þ	0.16	æ→e	0.27
u↔o	0.1	k↔g	0.24	pt→ft	0.31
y↔ö	0.1	y↔ö	0.24	g#→gg#	0.42
r#↔#	0.1	æ↔e	0.30	þer→n	0.43
þ↔t	0.1	å↔a	0.32	au→ö	0.44
þ↔th	0.1	þ↔t	0.32	th→þ	0.44
e#↔a#	0.1	z↔#	0.42	mp→m	0.45
#hv↔#v	0.1	r↔l	0.47	li→eli	0.45
f↔ff	0.1	r↔#	0.60	ghi→i	0.62

Figure 1: Example rules: hand-written (left), automatically derived character substitution (mid) and n-gram substitution (right). # marks word boundary.

a↔*e* is a cheaper substitution than *a*↔*h*.

From these alignments, we first extracted simple 1–1 substitution rules by collecting all character substitutions that were observed more than once in the data set. This resulted in 106 rules – a sample is found in Fig. 1 (mid). The costs are taken directly from the iterative Levenshtein distance alignment procedure. These 1–1 substitutions can only be used to match words of equal length. We thus extracted a more comprehensive set of substitutions by collecting all uni-, bi- and tri-gram alignments occurring more than once in the dataset. Because an n-gram in an alignment may contain ϵ-s, we not only create n–n rules, but also n–m rules. The cost associated with a substitution is $-\log p(\mathrm{RHS}|\mathrm{LHS})$, scaled down by a factor 10 to bring it in line with the costs of the other rule sets. Conditional probabilities were estimated using simple additive smoothing. Some examples of the resulting 6,045 rules are given in Fig. 1 (right).

3.3 Pilot Experiments

To get an impression of the coverage and accuracy of our approximate match approach, we applied our system to a small test collection consisting of a fragment of law text (Scania Law, Holm B 76) and a fragment of bible text (Marcus Gospel, 1526), together 249 tokens and 168 types based on string identity. The test corpus is too small to support more than a very coarse impression of the accuracy of our approach. However, the pilot evaluation is intended to give insight into the potential and challenges of our approach. We are planning a larger corpus to facilitate future development and allow more thorough evaluation. Because of the spelling variation, the manual annotation of the link-up is non-trivial.

	Match		Correct		Top 3		Top 10	
Method	#	%	#	%	#	%	#	%
exact match	70	*28*	55	*22*				
manual rules	147	*59*	118	*47*	122	*49*	122	*49*
auto char	197	*79*	116	*47*	131	*53*	134	*54*
auto n-gram	240	*96*	154	*62*	195	*78*	208	*84*

Table 1: Results of the pilot evaluation on 249 tokens.

For each word in the text, we checked whether the algorithm finds a match in one of the dictionaries, and whether this match is correct. We count a match as correct if the intended meaning of the word in its context is among the entries selected from the dictionary. We do not distinguish between homographs at this point, i.e., words with different sense or part-of-speech tag. We also count entries that refer to a correct entry as correct. For instance, we consider **ær.X**[4] in the Schlyter dictionary a correct match for *ær* 'is', as it refers to **vara.V**, which is the main entry for the verb *to be*. Such references are rare, though. Finally, we also applied a relaxed notion of correctness, using an oracle, where we consider a word as correctly linked if it is among the top three or top ten candidates. Note that we use the term *coverage* for both overall matches and correct matches. We avoid using the term *recall*, as we do not know how many words have a correct entry in the dictionary to begin with.

The results of the evaluation are in Table 1. As we can see, less than a third of the tokens in the texts had an exact match in one of the three dictionaries, and only counting correct exact matches, coverage is 22%, giving a precision of 79% (=55/70).[5] The manual rules roughly double this coverage, both in terms of overall matches and correct matches. Almost half of the tokens in the texts are now matched against the correct lexical entry. Note that the top 3 and top 10 oracles do

[4]Boldface indicates lexical entries.The following capital indicates part-of-speech as given by the dictionary, where X means unspecified.

[5]Considering that the used lexica are partly built upon these corpora, this is a low number. However, we match inflected forms against dictionary entries that are mostly in their base form.

not really increase the number of correct matches, as the rules only allow a very restricted amount of variation. Precision of the manual rules remains at 80% ($= 118/147$).

The automatically extracted character mapping fails to improve upon the correct match score of the manual rules, and precision drops to 59% ($= 116/197$). These rules thus create many misleading matches, compared to the manual rules. Finally, the automatically extracted n-gram rules find a match for almost all words in the text, and moreover, retrieve the greatest number of correct matches at 62% of all tokens. This, however, comes at a cost of low precision: 64% ($= 154/240$). When using the top *k* oracles the precision for the automatically extracted n-gram rules is instead encouragingly high at 81% ($= 195/240$) and 87% ($= 208/240$), respectively.

3.4 Evaluation and Future Work

One of the reasons that the automatically extracted n-gram method has such high overall coverage is that the substitution rules not only capture spelling variation, but inflection, too. For instance, for the inflected *öknen* 'desert.DEF.SG' we find the correct match **ökn.N**, a match we would have otherwise missed even though the spelling convention happens to be the same between source and target. In future work, we intend to treat morphology more systematically by incorporating the computational morphology mentioned in Section 2. For instance, we could use the morphology to lemmatize the text, after which the lemmata are linked-up to the dictionaries using our edit distance approach.

The top *k* oracles for the automatically extracted n-gram rules combine high coverage with high precision. We hope to be able to capitalize on this potential in the future by using context information to model the oracle. For instance, for *stijghar* in the sentence *görer hans stijghar retta* 'straighten his paths', we find the verb **stigha.V** 'ascend, walk' as a closer match than the noun **stigher.N** 'path'. Looking at the context, however, we may be able to guess that the nominal entry is more likely to be correct (see Wilcox-O'Hearn et al. (2008), Jurish (2010b) for the use of word level context in similar selection tasks).

Another avenue for research lies in the anagram hash filter. At the moment, we consider one million operations on the anagram hash, submitting on average just under 900 candidate matches for exact edit distance calculation. However, since we have over 6,000 substitution rules, many of which may apply for a particular word, even one million hash variations may only represent a fraction of the search space. When an even smaller number of operations is considered – attractive from a processing effort perspective – the filter removes many valid alternatives, either resulting in suboptimal top candidates or empty candidate lists. For instance, *cristindom*–**kristindomber.N** 'christianity, baptism' is found by the top 3 oracle in the 1 million operations, but not in the 10k filter. Such incorrectly filtered source-target pairs are typically long and either allow many different operations or require repeated applications of the same cheap rules. Resolving this mismatch between the prefilter step and the exact distance calculation is part of ongoing work.

We found that, in addition to linking the tokens to entries in the lexica, the entries need to be linked between the lexica. There is overlap between the dictionaries, but the different dictionaries give different information. For instance, ær 'is', mentioned above, is spelled *ær* in the law text, but *är* in the bible text. Only the Schlyter entry **ær.X** links this inflected form to **vara.V**. The corresponding entries in the Söderwall dictionaries, **är.X**, give only noun, conjunction and pronoun uses. Consolidating the dictionaries will increase the amount of information available.

Finally, the evaluation revealed the necessity to be able to handle multi-word tokens. For example, the text contains the expression *köpæ iorth* 'bought land'. While the word *iorth* is correctly matched in the lexicon, the lexicon entries found for *köpæ* are incorrect, as they refer to the verb buying or the people involved in buying. The entry **köpe iorþ.N** does appear in two of the lexica but cannot presently be found as we only consider graphic words. There are also cases where we need to split a graphic word to match it against a dictionary. For instance, the parts of the compound *villhonnugh* 'wild honey' can separately be matched against **vilder.AV** 'wild' and **hunagh.N** 'honey'.

Splitting and merging compounds and matching them against the dictionary can be readily integrated by allowing substitution rules to contain

graphic word boundaries and considering multiple source tokens at one time (Reynaert, 2011). An effective way of doing separate matching as needed in the case of *vill-honnugh* remains a question for future work.

4 Conclusions

We are working towards automatic annotation of about 3 million words of historical Swedish texts. As a first step, we are developing a module to handle spelling variation. Three sets of approximate matching rules were used, one with hand-crafted substitutions, and two with substitutions automatically extracted from alternative entries in a lexicon. We presented a pilot evaluation by matching words in two short texts to our historical lexica. While the automatically created rules find more matches, the manual rules have higher precision. Future work includes improving the anagram hash filter, incorporating Old Swedish morphology, handling multi-word units, and exploiting context information to improve precision.

Acknowledgements

The research presented here is carried out in the context of the Centre for Language Technology of the University of Gothenburg and Chalmers University of Technology.

References

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the RANLP Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42. Hissar, Bulgaria.

Lars Borin and Markus Forsberg. 2008. Something old, something new: A computational morphological description of Old Swedish. In *Proceedings of the LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 9–16, Marrakech, Marocco. ELRA.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology. In Nivre, Dahllöf, and Megyesi, editors, *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, volume 7 of *Studia Linguistica Upsaliensia*, pages 21–32. Uppsala University, Department of Linguistics and Philology, Uppsala, Sweden.

Lars Borin, Markus Forsberg, and Dimitrios Kokkinakis. 2010. Diabase: Towards a diachronic BLARK in support of historical studies. In Calzolari, Choukri, Maegaard, Mariani, Odijk, Piperidis, Rosner, and Tapias, editors, *Proceedings of the LREC 2010 workshop on Semantic relations. Theory and Applications*, Valletta, Malta. ELRA.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. In Calzolari, Choukri, Declerck, Doğan, Maegaard, Mariani, Odijk, and Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA.

Anders Fredrik Dalin. 1853/1855. *Ordbok öfver svenska språket*, volume I–II. Stockholm, Sweden.

Bryan Jurish. 2010a. Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 72–77, Uppsala, Sweden. Association for Computational Linguistics.

Bryan Jurish. 2010b. More than Words: Using Token Context to Improve Canonicalization of Historical German. *Journal for Language Technology and Computational Linguistics*, 25(1):23–40.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2012. Parsing the past - identification of verb constructions in historical text. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.*, Avignon, France.

Martin Reynaert. 2011. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14:173–187. 10.1007/s10032-010-0133-5.

Carl Johan Schlyter. 1887. *Ordbok till Samlingen af Sweriges Gamla Lagar*, volume 13 of *Saml. af Sweriges Gamla Lagar*. Lund, Sweden.

Knut Fredrik Söderwall. 1884/1953. *Ordbok Öfver svenska medeltids-språket. Ordbok Öfver svenska medeltids-språket. Supplement*. Lund, Sweden.

Martijn Wieling, Jelena Prokić, and John Nerbonne. 2009. Evaluating the pairwise string alignment of pronunciations. In Borin and Lendvai, editors, *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R 2009) Workshop at the 12th Meeting of the EACL*, pages 26–34.

Amber Wilcox-O'Hearn, Graeme Hirst, and Alexander Budanitsky. 2008. Real-word spelling correction with trigrams: A reconsideration of the mays, damerau, and mercer model. In Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 4919 of *LNCS*, pages 605–616. Springer.