

Computer-aided morphology expansion for Old Swedish

Yvonne Adesam,¹ Malin Ahlberg,¹ Peter Andersson,¹
Gerlof Bouma,¹ Markus Forsberg,¹ Mans Hulden²

¹Språkbanken, Department of Swedish, University of Gothenburg, Sweden

²Department of Modern Languages, University of Helsinki, Finland

yvonne.adesam@gu.se, malin.ahlberg@gu.se, peter.andersson@gu.se
gerlof.bouma@gu.se, markus.forsberg@gu.se, mans.hulden@helsinki.fi

Abstract

In this paper we describe and evaluate a tool for paradigm induction and lexicon extraction that has been applied to Old Swedish. The tool is semi-supervised and uses a small seed lexicon and unannotated corpora to derive full inflection tables for input lemmata. In the work presented here, the tool has been modified to deal with the rich spelling variation found in Old Swedish texts. We also present some initial experiments, which are the first steps towards creating a large-scale morphology for Old Swedish.

Keywords: Old Swedish, paradigm induction, lexicon extraction

1. Introduction

Language technology for historical texts has been the target of much recent interest. In some cases, the historical texts processed are similar enough in language to contemporary texts to allow the reuse of automatic methods developed for contemporary material. Often, however, the differences between historical and contemporary materials are substantial. That is the case, for instance, between Old Swedish (13–16th c.) and contemporary Swedish.

Old Swedish, compared to contemporary Swedish, is both morphologically and grammatically different. E.g., it has more cases, verb congruence, a different word order, namely OV (object-verb), verb-final subordinate clauses, and possibly missing subject. In addition, there is no single orthographic standard for written text, which results in a wide variety of spellings (even for the same word within a paragraph), and a variety of boundary marking strategies, which even make it complicated for an Old Swedish scholar to decide what is a word, and where one sentence ends and another begins.

As an example, consider the inflection table of the noun *fisker* ‘fish’ in Table 1. The inflection table captures the rich suffix variation; stem variation is assumed to be dealt with elsewhere. Compared with the contemporary Swedish word for ‘fish’, *fisk*, it has four times as many word forms. Furthermore, as reference, we have also included the inflection table of *fisker* as presented in traditional grammatical descriptions of Old Swedish (Wessén, 1969; Wessén, 1971; Wessén, 1965; Noreen, 1904; Pettersson, 2005), which contain fewer word forms due to normalization.

In this paper, we describe the application of an automatic method for morphology expansion to data from Old Swedish. The aim is to extend the manually created computational morphology of Old Swedish described in Borin and Forsberg (2008) with entries found in existing digitized dictionaries for Old Swedish (Schlyter, 1877; Söderwall, 1884–1973). Underlying our interactive tool are algorithms for semi-supervised paradigm induction and lexicon extraction, which have been shown to be effective on contemporary languages (Ahlberg et al., to appear). Their application to

lemma	fisker				traditional normalized form
PoS	nn				
gender	m				
num	def	case	word form		
sg	indef	nom	<i>fisker</i>	<i>fisker</i>	
sg	indef	gen	<i>fisks</i>	<i>fisks</i>	
sg	indef	dat	<i>fiski, fiske, fisk</i>	<i>fiski, fisk</i>	
sg	indef	acc	<i>fisk</i>	<i>fisk</i>	
pl	indef	nom	<i>fiska(r), fiskæ(r)</i>	<i>fiska(r)</i>	
pl	indef	gen	<i>fiska, fiskæ</i>	<i>fiska</i>	
pl	indef	dat	<i>fiskum, fiskom</i>	<i>fiskum</i>	
pl	indef	acc	<i>fiska, fiskæ</i>	<i>fiska</i>	
sg	def	nom	<i>fiskrin</i>	<i>fiskrin</i>	
sg	def	gen	<i>fisksins</i>	<i>fisksins</i>	
sg	def	dat	<i>fiskinum, fisk(e)num</i>	<i>fiskinum</i>	
sg	def	acc	<i>fiskin</i>	<i>fiskin</i>	
pl	def	nom	<i>fiskani(r), fiskæni(r)</i>	<i>fiskani(r)</i>	
pl	def	gen	<i>fiskanna, fiskænna</i>	<i>fiskanna</i>	
pl	def	dat	<i>fiskumin, fiskomin</i>	<i>fiskumin</i>	
pl	def	acc	<i>fiskana, fiskæna</i>	<i>fiskana</i>	

Table 1: The inflection table of *fisker* ‘fish’

Old Swedish, however, presents a number of challenges that stem from both the relative lack of resources and the high variability in the existing resources.

2. Computer-aided morphology expansion

By morphology expansion, we refer to the task of building a broad-coverage morphological description of a language from a smaller seed lexicon. The morphology expansion method described in Ahlberg et al. (to appear) operates in two steps. In the first step, compact paradigm descriptions are induced from a seed of inflection tables. In the second step, new instances of these paradigms are suggested to the expert user, who can then decide whether to incorporate these into the morphology.

2.1. Paradigm induction

In the first step, the tool induces abstract paradigm descriptions from known full inflection tables. The central

principle in the induction algorithm is the identification of the longest common subsequence (LCS) shared by all forms in an inflection table. This divides each cell in the table into inflected and non-inflected parts. By abstracting away from the non-inflected parts, we arrive at a paradigm description in the form of string patterns.

The algorithm is illustrated in Figure 1 for English *ring*~*rang*~*rung* and *swim*~*swam*~*swum*. The respective LCSs, that is, the non-inflected parts, are *rng* and *swm*, which leaves the alternation *i*~*a*~*u* as the inflection in each table. The two input tables can therefore be regarded as instances of the same paradigm.

The induced paradigms encode all variation within the inflection tables, such as affixation, phonological alternation and orthographic change, in a compact representation.

2.2. Lexicon extraction

In the second step, the tool expands the lexicon consisting of part-of-speech tagged lemmata by automatically assigning induced paradigms to them using corpus data. Given a lemma and a paradigm, the tool first decides whether they are compatible, that is, whether the lemma matches the pattern of the paradigm’s base form cell. If compatible, a hypothesized full inflection table is generated and a confidence score for that table is calculated. The score is a product of the following two measures:

analogy: the amount of overlap between suffixes of the input lemma and the lemmata known to belong to the paradigm, in terms of length as well as number of overlapping suffixes;

frequency: the combined corpus frequency of the hypothesized inflection table, spread out over unique forms:

$$\sum_{w \in \text{set}(\text{Forms})} \log(\text{count}(w) + 1).$$

After scoring all possible lemma-paradigm pairs, each lemma is assigned to the highest-scoring paradigm. The number of previously seen instances of a paradigm is used as a final tie breaker. In an alternative setup, we focus on a single paradigm and present the user with the highest-scoring lemma for that paradigm.

3. Handling spelling variation

Our paradigm induction and lexicon extraction methods are largely language-independent and can be applied to any part of the (paradigmatic) morphological system of a language. The analogy measure described above assumes

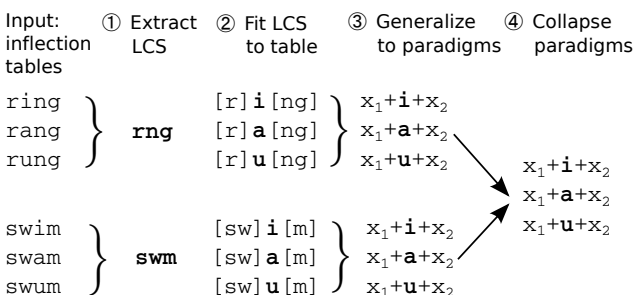


Figure 1: The paradigm induction procedure

that the language is suffix-inflecting. Old Swedish should be a good fit for our method, though it presents challenges with regard to the extensive variation in spelling.

There are several reasons for a high degree of variation in the Old Swedish corpus material. First, the material spans nearly three centuries and comes from a variety of geographic locations in a time without a common orthographic standard. Thus, it exhibits a wide variety of spellings (even within a single paragraph) and a variety of boundary-marking strategies. Second, the Old Swedish period was characterized by grammatical change in inflectional complexity (case, gender, verb congruence, etc.). Finally, we are working with editions upon editions of the original manuscripts.

The more systematic types of variation we have observed include (near-)equivalence between characters (*a*~*ä*, *æ*~*ö*~*ø*, *u*~*v*~*w*, *o*~*u*, *i*~*j*), pronunciation changes/differences or variation in feature marking (lowering: *manni*~*manne* ‘person.OBL’, voicing: *gfin*~*gvin* ‘given’, length: *hor*~*hoor* ‘adultery’, fronting: *magher*~*mogher*~*mågh* ‘son in law’, strengthening: *efter*~*epter* ‘after’), and more general multi-character correspondences (*ghi*~*i*~*ε*: *flyghia*~*flyia*~*flya* ‘flee’, *þ*~*dh*~*th*~*d*: *maþer*~*madher*~*mather*~*mader* ‘person’).

The lemma lists used in the lexicon extraction step come from the available Old Swedish dictionaries. Although dictionary entries are normalized forms, they do not provide us with a particularly regular spelling system. The two dictionaries adhere to different orthographic criteria, and even within one dictionary we find irregularities due to the application of multiple, conflicting normalization principles, such as etymology, attestation, frequency, and (constructed) pronunciation (Djäv, 2009).

Spelling variation is mainly a problem in calculating the frequency part of the confidence score, as exact matching of a form against the corpus underestimates the occurrence frequencies. We handle this in two ways. First, the Old Swedish morphology of Borin and Forsberg (2008) treats spelling variation in inflectional suffixes as (free) allomorphy. The induced paradigms and the hypothesized full inflection tables thus include suffix spelling variants. Second, we use fuzzy matching to handle spelling variation in the rest of the forms. The fuzzy matching module (Adesam et al., 2012) contains weighted rewrite rules that capture variations like those exemplified above. We consider words with a low rewrite cost to be spelling variants, and thus to constitute matches for a hypothesized form. For instance, the observed forms *kunungher* and *konunger* can count as corpus matches for a (correctly) hypothesized form *kununger*, as they only require a cheap deletion (*h*→*ε*) and a cheap substitution (*o*→*u*), respectively. Since suffix variants are generated from the paradigm, no additional rewrites in the suffix are allowed.

4. Experiments

4.1. Experimental setup

We applied our system to Old Swedish nouns to assess its viability and effectiveness as a tool for morphological lexicon building in a low-resource, high-variation language. We used inflection tables generated by the morphology of

	correct	incorrect	unclear
Top 20	18	2	0
Top 100	54	38	8
p_fisker @100	14	3	1
p_gata @100	14	8	3
p_biti @100	12	14	4

Table 2: Results of Experiment 1

Borin and Forsberg (2008) for step one, paradigm induction. From 3 500 nominal inflection tables, generated using the 38 noun paradigms of Borin and Forsberg (2008), we induced 46 paradigms. The lemma list for step two, lexicon extraction, consists of all nouns listed in the Old Swedish dictionaries, minus those that were already present in the paradigm induction input – a total of $\sim 15k$ lemmata. Finally, 2.5 million tokens of Old Swedish text from Fornsvenska Textbanken¹ were used as reference corpus.

We consider two different scenarios for our experiments: in the first experiment, a set of paradigms are given, and the task is to decide which word belongs to which paradigm. In the second experiment, only one paradigm is given, and the task is to rank the words that are most likely to belong to this paradigm. In the first case, paradigms compete against each other and only those yielding the highest confidence score for a word will be suggested to the expert. While this competition between the different paradigms may improve the result, the second experiment corresponds to a more common scenario: the expert user has finished editing a new paradigm and wants to find words that are likely to match its description.

Ultimately, we envisage an iterative use of the tool: the expert user adds new paradigms, selects additions suggested by the system, goes back to editing or adding paradigms, asks the system for more suggestions, etc. The user will never consider all suggestions the system makes, only those with high confidence scores.

4.2. Evaluation

In the first experiment, where all paradigms are considered for each input word, but only the best is chosen, we evaluated the 100 highest-ranked word-paradigm pairs. The results are listed in Table 2. From a total of 46 competing paradigms, 17 are selected at least once in this top 100. Three of them – the masculine weak paradigm **p_biti**, the feminine weak **p_gata** and the masculine strong **p_fisker** – together constitute 73 out of 100 suggestions.

Of the top 100 suggestions, 54 are correct and 8 unclear, where ‘unclear’ means that the expert user was unable to judge if the paradigm assignment was correct or not. The accuracy of the top 20 is significantly higher, with 18 correct assignments, which indicates that the confidence score is effective. We also note differences in accuracy per paradigm in the top 100, ranging from less than half correct for **p_biti** to a substantial majority correct for **p_fisker**.

In the second experiment, we evaluated the top 25 lemmata of one paradigm at a time, disregarding the confidence

	correct	incorrect	unclear
p_gata	13	11	1
p_kyrkia	13	6	6

Table 3: Results of Experiment 2

scores for the competing paradigms. In Table 3 we present the results for **p_gata** and **p_kyrkia**. The former was a common paradigm in the results of experiment 1, while the latter has many instances in the computational morphology. For both paradigms, about half the suggestions were judged correct. However, **p_kyrkia** has a larger number of *unclear* judgements.

4.3. Discussion

The accuracies reported in the previous section are good enough for the tool to be useful in an interactive setting, especially if the expert user is able to restrict himself to the top results. However, compared to the results with several contemporary languages reported in Ahlberg et al. (to appear), the Old Swedish results represent a considerable drop in performance.

A common source of error in experiment 1 is the mix-up of similar paradigms. For instance, the low accuracy for the masculine **p_biti** is due to incorrect assignments from the neuter paradigm **p_æpli**. These have a similar base form but different oblique forms. Due to data sparseness, these differences may not be visible in the frequency rating. This suggests that additional information provided by an expert user—e.g., that a subset of the word forms is crucial for making the correct distinctions—should be incorporated into the confidence score in the future. Such information may, in fact, have already been encoded in the dictionaries. For instance, *riki* ‘country’ should never be assigned to the masculine paradigm **p_biti**, since the dictionaries dictate that it is a neuter noun, i.e., **p_biti** should be rejected in favor of **p_æpli**. In fact, the dictionaries also provide the genitive definite *rikisins*, which is not a form compatible with **p_biti**. Including such information would be an interesting extension of this work.

The corpus data lacks part-of-speech tags, which means that high-frequency word forms of another part of speech than the current lemma may inflate the score for a particular paradigm. A part-of-speech tagged corpus would obviously alleviate this problem, but no such corpus for Old Swedish currently exists.

Contrary to our experience with modern Swedish, having competing paradigms for a lemma did very little to improve accuracy. The problem stems from two factors: (1) the rich form variation occurring in the input inflection tables, resulting in a high degree of overlap between paradigms; (2) the data sparseness. Together they mostly reduce the competition to just a random selection.

Finally, a general difficulty in both experiments is the fact that many lemmata are very rare or even unattested in the available resources. This lack of empirical evidence for the correct inflection of a word is reflected in the high number of unclear judgements for certain paradigms.

¹<http://project2.sol.lu.se/fornsvenska>

5. Future work

There are several immediate directions to increase the usefulness of the system. For example, lemmata for which only the base form can be attested, or which refer to more common base forms, could be ignored, to avoid making predictions on scanty or no empirical data. Furthermore, the dictionary entries in many cases contain more word form examples and grammatical information than just the lemma and the part of speech, which can be used in the lexicon extraction process.

For the spelling variation, a cruder normalization has been successfully applied in other applications (Bouma and Adesam, 2013).

Moreover, an ongoing effort is to develop a graphical tool, named *Morfologilabbet*² 'the morphology lab', that incorporates the methods presented in this article. A screenshot showing the current state of the tool is displayed in Figure 2. Without going into details, the result of the lexicon extraction is shown on the left, and the inflection table of *dagher* 'day' on the right, here assigned the paradigm of *fisker* 'fish'. Furthermore, the word forms in the inflection table are annotated with frequency. A more detailed description of the tool will be published in the near future.

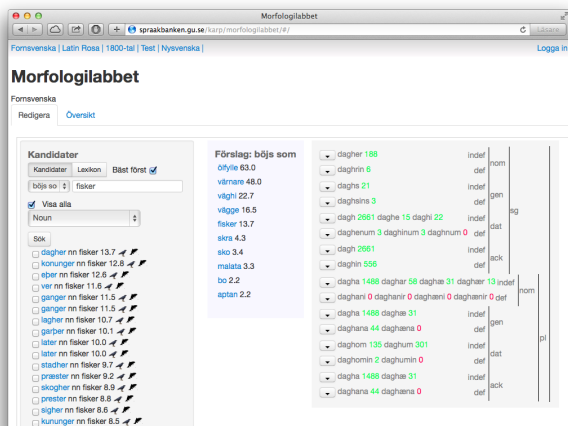


Figure 2: *Morfologilabbet* 'the morphology lab'

6. Conclusions

We have described the application of a method for paradigm induction and lexicon extraction to Old Swedish data, which presents a challenge to any computational processing as a low-resource, high variability language. While the tool's accuracy is considerably lower than what is achievable for contemporary languages, we are convinced that the system is valuable for morphology development in one of the interactive settings described here. The pilot experiment has shown a need to incorporate more information into the lexicon extraction process. We are currently addressing this in the ongoing development of the tool.

Acknowledgements

The research presented here was supported by the Swedish Research Council (the projects *Towards a knowledge-based cul-*

turoemics, dnr 2012-5738, and *Swedish Framenet++*, dnr 2010-6013), the Marcus and Amalia Wallenberg Foundation (*MAPiR* project, MAW 2012.0146), the University of Gothenburg through its support of the Centre for Language Technology and its support of Språkbanken, and the Academy of Finland under the grant agreement 258373, *Machine learning of rules in natural language morphology and phonology*.

7. References

- Adesam, Yvonne, Ahlberg, Malin, and Bouma, Gerlof. (2012). bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa... towards lexical link-up for a corpus of old swedish. In *Proceedings of the LTHist workshop at Konvens*.
- Ahlberg, Malin, Forsberg, Markus, and Hulden, Mans. (to appear). Semi-supervised learning of morphological paradigms and lexicons. In *European Chapter of the Association for Computational Linguistics (EACL 2014)*.
- Borin, Lars and Forsberg, Markus. (2008). Something old, something new: A computational morphological description of Old Swedish. In *LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 9–16.
- Bouma, Gerlof and Adesam, Yvonne. (2013). Experiments on sentence segmentation in Old Swedish editions. In *NEALT Proceedings Series*, volume 18.
- Djärv, Ulrika. (2009). *Fornsvenskans lexikala kodifiering i Söderwalls medeltidsordbok*. Ph.D. thesis, Uppsala University.
- Noreen, Adolf. (1904). *Altschwedische Grammatik*. Halle.
- Pettersson, Gertrud. (2005). *Svenska språket under sjuhundra år*. Studentlitteratur, Lund, Sweden.
- Schlyter, Carl Johan. (1877). *Ordbok till Samlingen af Sweriges Gamla Lagar*, volume 13 of *Saml. af Sweriges Gamla Lagar*. Lund, Sweden.
- Söderwall, Knut Fredrik. (1884–1973). *Ordbok öfver svenska medeltids-språket*, volume I–III, IV–V (supplement). Lund, Sweden.
- Wessén, Elias. (1965). *Svensk språkhistoria: Grundlinjer till en historisk syntax*. Stockholm, Sweden.
- Wessén, Elias. (1969). *Svensk språkhistoria: Ljudlära och ordböjningslära*. Stockholm, Sweden.
- Wessén, Elias. (1971). *Svensk språkhistoria: Ordböjningslära*. Stockholm, Sweden.

²<http://spraakbanken.gu.se/karp/morfologilabbet>