NEALT Proceedings
Northern European Association for Language Technology

Proceedings of the Workshop on
Computational Historical Linguistics

NODALIDA 2013

May 22-24, 2013 • Oslo, Norway

Linköping Electronic Conference Proceedings

Proceedings of the workshop on

computational historical linguistics

at NODALIDA 2013

edited by

Þórhallur Eyþórsson
Lars Borin
Dag Haug
Eiríkur Rögnvaldsson

NEALT Proceedings Series 18 • ISBN 978-91-7519-587-2
Linköping Electronic Conference Proceedings 87
ISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2013

Preface

Recent years have seen a surge of interest in the application of computational methods to problems in historical linguistics. To date, much of this work has been based on the application of simple similarity measures to short lists of lexical items or grammatical features for achieving large-scale genetic grouping of languages. While highly publicized and demonstrably useful, such approaches are inherently limited both by the narrow range of linguistic features examined and the low-level processing methods used.

At the same time, language technology for dealing with modern languages has developed apace, with automatic language tools now achieving a degree of accuracy that has enabled both popular online services such as Google translate and the rapid accumulation of linguistically annotated monolingual and multilingual corpora for many languages. Much less has been done on historical texts: there is little commercial interest in these language varieties, there is often limited amounts of data (making purely data-driven annotation approaches unfeasible), and they are less well-behaved than modern print corpora, due to lack of standardization on all linguistic levels, starting with orthography. Digitized older texts also often suffer from OCR errors.

The basic premise of the workshop is that historical linguistics can benefit greatly from having access to historical and diachronic corpora with rich linguistic annotations, but this is a field where researchers have barely scratched the surface of what is possible. However, because of the nature of the material and of the research questions, interesting questions of theory and method arise in connection with this work, which often are relevant to work on modern data as well (e.g., linguistic variation in spoken language or in web genres). The workshop aimed at providing a forum where these questions can be discussed. The target audience of the workshop were researchers – linguists and computational linguists – involved in the creation and utilization of richly annotated historical and diachronic text corpora, in the context of historical-comparative (diachronic, genetic) linguistic research.

We invited papers presenting original research relating to computational historical linguistics, on topics such as:

- theoretical and methodological aspects of automatic annotation for historical linguistic research, e.g.:
  - the influence and significance of annotation errors
  - which kinds of annotation are needed and useful for historical linguistics
  - how to deal with variation and multilinguality
  - annotation transfer between diachronic language stages or between languages
  - issues of standardization, interoperability and data sharing
- innovative user interfaces for computational historical linguistics (including search and visualization solutions)
- design of optimal annotation workflows with manual and automatic components for creating historical and diachronic corpora
- linguistic processing of annotated historical and diachronic corpora for historical linguistic research, e.g.:
  - methods for tracking change in vocabulary and grammar in diachronic corpora
  - grammar extraction and comparison on historical and diachronic treebanks

Six submissions were accepted for presentation at the workshop and inclusion in this proceedings volume after a thorough review procedure and subsequent revision by the authors of the papers. Each submission was reviewed by three (anonymous) members of the program committee:

- Yvonne Adesam (University of Gothenburg)
- David Bamman (Carnegie Mellon University)
- Lars Borin (University of Gothenburg)
- Gerlof Bouma (University of Gothenburg)
- Stefanie Dipper (Ruhr-Universität Bochum)
- Michael Dunn (MPI for Psycholinguistics, Nijmegen)
- Þórhallur Eyþórsson (University of Iceland)
- Markus Forsberg (University of Gothenburg)
- Dag Haug (University of Oslo)
- Seth Kulick (Linguistic Data Consortium)
- Hrafn Loftsson (Reykjavik University)
- Marco Passarotti (Catholic University of the Sacred Heart, Milan)
- Michael Piotrowski (Leibniz Institute of European History)
- Eiríkur Rögnvaldsson (University of Iceland)

The workshop featured two invited speakers: **Seth Kulick** (Linguistic Data Consortium), who gave a presentation with the title *Treebank analysis using derivation trees*, and **Michael Piotrowski** (Leibniz Institute of European History), who talked about *Historical NLP and the Digital Humanities*.

The workshop organizers:
Þórhallur Eyþórsson
Lars Borin
Dag Haug
Eiríkur Rögnvaldsson

WS website: `http://spraakbanken.gu.se/swe/nodalida-chl-ws-2013`

Acknowledgements: Financial support for the organization of the workshop has come in part from NOS HS (the exploratory workshop grant *Diachonic Syntax Corpus (DiaSynCorp)*).

Contents

Towards automatic tracking of lexical change: linking historical lexical resources

Malin Ahlberg, Peter Andersson

Språkbanken / Department of Swedish
University of Gothenburg

malin.ahlberg@gu.se, peter.andersson@gu.se

ABSTRACT

In the field of historical linguistics, large-scale corpora studies are a key component in identifying phenomena such as language variation and language change. Manually performed corpora studies are very time consuming and may obscure interesting changes in the sense that phenomena that are not being specifically searched for easily are overlooked. During the last couple of years the potential of language technology tools has been put forward in relation to historical linguistic research. This paper is based on an experiment of linking up several lexical resources in Swedish, which together reflect a vocabulary from Old Swedish to Contemporary Swedish. The link-up aims at identifying potential lexical change such as cases of grammaticalization and it may further be of use in other language technology applications. In our case study we are linking lemmas together with part-of-speech information given in each entry for all the lexical resources. This paper describes our first results, where we focus on the cases when information about word class differs in one of the resources. In future studies it is necessary and desirable to include more digitalized lexicon resources and confirm the analyses with corpora research. Still, the current result already shows some interesting cases of semantic change and grammaticalization. Changes in the content word system such as generalization and specialization of meaning are also exemplified in our data. Even though the links sometimes show errors that at first sight lead us towards a wrong conclusion we believe that methods like the one used here may be very fruitful to future research to reach more efficiency in historical linguistics research.

KEYWORDS: lexical linking, historical linguistics, language change.

Malin Ahlberg and Peter Andersson 2013. Towards automatic tracking of lexical change: Linking historical lexical resources. *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013.* NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87: 1–10.

1 Introduction

During the last couple of years there has been an increased interest in the potential of language technology tools for other disciplines than modern linguistic research. Recent work (see e.g. Pumfrey and Mariani, 2012) has shown that large-scale corpus-based historical linguistics is a fruitful approach. An obvious area of collaboration between historical linguistics and language technology research is language change, such as grammaticalization. The work described in this paper aims at identifying some aspects of lexical change starting from an experiment on a link-up of a large bank of lexical resources.

The purpose of the article is twofold: first, we describe a link-up of historical lexica to modern resources, where existing lexica is integrated into a resource functioning as a dictionary between Old and Contemporary Swedish, suitable for language technology applications. It may be used to increase the readability of old texts, in information retrieval for historical corpora and for annotating and analysing these. Second, we give a case study of how large-scale comparisons of lexica from different time periods may help identifying and conceptualizing language change. Our intention is to give an example of how more efficient methods, in this case a set of automatically created links between historical and modern lexical units, can be used to support the formation of new research questions. Even if this approach generates some links without any apparent relation and obscures or misses others, the method may identify lexical change that we never have discovered with manual work. The case study in this paper examines the lemmas that exist in both historical and contemporary lexica and compares them with respect to which word classes they are assigned by different dictionaries.

Our central resources are five lexical resources available at Språkbanken[1]: Söderwall (1884), its supplement(Söderwall, 1953) and Schlyter (1887) for Old Swedish, Dalin (1855) for Modern Swedish, Saldo(Borin et al., 2008) for Contemporary Swedish. The rough distinction between language stages is based on the frequently used subdivision summarized in (Andersson, 2007). Söderwall's and Schlyter's dictionaries reflect Old Swedish (henceforth OSwedish) up to about year 1526 with the bible translation of The New Testament as a landmark. The period thereafter and up to about year 1900 is usually subsumed as Modern Swedish, for younger language the term Contemporary Swedish is used. We refer to Modern and Contemporary Swedish as MCSwedish.

The historical lexica have been digitized and are all available in LMF[2] format. They contain totally 50 000 entries of OSwedish, but the lexica overlap and the actual number of separate lexical entries is assumed to be lower. On the other hand, most entries contain a number of different senses, among which the semantic difference may be quite big. Saldo, a semantic and morphological lexicon of Contemporary Swedish, contains 120 000 entries. It is the central component in the Swedish FrameNet++ (SweFN++) (Borin et al., 2010), in which a number of lexical resources are integrated. Our work connects the three historic dictionaries to SweFN++ by a simple matching procedure presented in Section 3. In Section 4 we give examples and discuss cases of semantic change and grammaticalization that are shown when comparing the results of the link-up.

[1] http://spraakbanken.gu.se
[2] http://www.lexicalmarkupframework.org

2 Background

The examples of link-ups and integrations between modern lexical resources are many. Babelnet (Navigli and Ponzetto, 2012) is a multilingual ontology, constructed from an automatic integration of Wikipedia and WordNet. de Melo and Weikum (2009) combine a number of wordnets with information from dictionaries and parallel corpora into the Universal WordNet. Another example is Uby (Gurevych et al., 2012) combining English and German lexical resources.

For historical lexica, the digital resources are more scarce. Gotscharek et al. (2011) present an interactive tool for constructing lexicons from corpora, which they use to produce a historical lexicon for German with 10 000 entries. In information retrieval for historical corpora, most research however normalizes the text to modern spelling, Koolen et al. (2006) present some different methods for this. Ernst-Gerlach and Fuhr (2007) on the other hand transform a given search query in modern language into a number of possible historical forms. Borin and Forsberg (2011) present a diachronic lexical resource, where Dalin is linked to Saldo. We will refer to this material as the Diapivot. Their work further includes the start on a link-up between the OSwedish lexica and Saldo, which constitutes a starting point of our work.

In the case study of this paper we mainly focus on identifying semantic changes and grammaticaliziation. Semantic change includes processes such as generalization and specialization hence broadening and narrowing of meaning. Grammaticalization is the well described unidirectional process of correlating semantic, morpho-syntactic and phonological developments, often resulting in change of word class (Hopper and Traugott, 2003; Andersson, 2007). Traugott (2001) defines grammaticalization as "the change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions or grammatical items develop new grammatical functions." Frequently cited examples are the development of auxiliaries from main verbs such as the change of the Swedish main verb *verka* 'act, work' to an auxiliary *verka* 'seem' or the adjective *bar* 'naked' which has developed to an adverb *bara* 'only' and further into a speech particle *ba* (Eriksson, 1992).

3 A (transitive) link-up from Old to Contemporary Swedish

The goal of the link-up discussed in this paper is to match the lemmas given in three OSwedish lexica to corresponding lemmas in modern lexica. The language has undergone much orthographical change and standardizations leading to a vast differences between Old and Contemporary Swedish. In order to find links between lemmas between which there is no longer a (visible) linguistic connection, we base the linking on information given within the entries. These often include a translation to Late Modern Swedish (see Figure 1), which we extract

äptirfinna vb	*återfinna. "haffuer ther til badhä hug och sinne thz skiutast iach ma idher effter finnä" Iv 1912 .*
värdhskylda vb	*förtjena. "hwat ey wärdhzskylladhe war syndh wphöghilse, sannelica ney Skrifter till uppbygg. 73."*
æpli nn	*äple. VG. lIII. 123; U. Kk. l7: 5. not. 45; ME. * St. Eds. 25; Thj. 6.*

Figure 1: Examples of lexical entries in the Old Swedish lexica

together with the lemma it self and its part-of-speech tag. The definitions are then compared to the lemmas in Dalin - if a match which in turn is linked to Saldo via the Diapivot is found, we create a link from the OSwedish lemma to the contemporary entry. Examples are shown in Figure 2a-b. The results are promising, we get 16 000 links of surprisingly good quality

	OSwedish		definition		Dalin		Saldo
a.	värdhskylda (vb)	→	förtjena	→	förtjena (vb)	→	förtjäna (vb)
b.	æpli (nn)	→	äple	→	äple (nn)	→	äpple (nn)
c.	aterhitta (vb)	→	återfinna	→			återfinna (vb)
d.	ivirbygning (nn)	→	överbyggnad	→			överbyggnad (nn)
e.	ighul (nn)	→	igelkott	→			igelkott (nn)
f.	sundrisker (adj)	→	söderländsk	⇝			sörländsk (adj)
g.	hvitklädder (adj)	→	hvitklädd	⇝			vitklädd (adj)
h.	tyghi (nn)	→	vitnesbörd	⇝			vittnesbörd (nn)
i.	kokiöt (nn)	→	kokött	⇝			*kokott (nn)

Figure 2: Links connecting OSwedish lemmas to Saldo

(see below). We further extend this link-up by allowing links to be created directly between OSwedish and Saldo, whenever there is no match in the Diapivot (see Figure 2c-e). Finally, the yet unlinked lemmas are matched against Saldo entries using spelling normalisation. Using techniques from Adesam et al. (2012), we extracted spelling variation rules from the Diapivot, capturing differences between Modern and Contemporary Swedish. Example of (correct and incorrect) matches found this way is found in Figure 2f-i. A fuzzy matching like this obviously introduces more errors to the link-up and methods to minimize these are part of our future work.

At this point, we only give a very limited pilot evaluation, based on 150 randomly chosen pairs of linked lemmas (see Figure 4). We mark a link as correct whenever the given Saldo entry corresponds to the definition in the OSwedish lexica. For the Diapivot-matching and the fuzzy matching we also accept broad matches, which, for the first case, means that a lemma might link to a hypernym and in the latter means that adverbs may link to adjectives, cf. *plötsligt* 'suddenly' → *plötslig* 'sudden'. The erroneous links correspond to cases when the spelling normalisation were too generous, when the extraction of a definition failed to identify the relevant parts of information, or when the linking could not identify the correct modern lemma from a set of homographs. Figure 3 shows the number of created links per word class. Black marks the links created by using the Diapivot, gray the links created directly from Old to Contemporary Swedish, and white the links created using fuzzy matching.

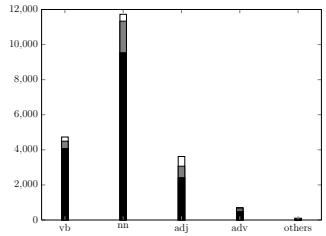

Figure 3: Number of linked lemmas per word class.

	Diapivot	Direct	Fuzzy
Correct	86%	86%	82%
Erroneous	14%	14 %	18%

Figure 4: Correctness of the different methods

3.1 Improving the link-up

An important step to increase the coverage as well as the accuracy of the link-up is to improve the quality of the digitized lexica. In their present form, inconsistent mark-up sometimes hinders us from identifying the correct definition. We are also working on refining the method for extracting the most relevant parts of a definition, to avoid mistaking lexical terms and abbreviations for being part of the real definition. A third improvement is to identify entries of the OSwedish lexica that are defined by referring to another entry within the same lexicon. This is quite common, and by interlinking these pairs and letting them share their connections to Saldo we could increase the coverage of the link-up. In Borin and Forsberg (2011), some links between Saldo and Dalin are created using a basic analysis of unknown lemmas in Dalin. For instance, unknown compounds may be linked by their head word only. This is the source of the broad matches mentioned above, and the approach would be helpful also in our case.

We further need to consider how entries with multiple matches in Saldo should be handled, and to what degree the linking should be carried out on a level of senses rather than lemmas. The semantic difference between two senses of one lemma in the OSwedish lexica may be rather big, cf. the lemma *leka* which apart from *spela* 'play' is defined as both *strida* 'fight' and *roa sig* 'have fun'. For comparison, the corresponding modern word *spela* is divided into six separate entries in Saldo. Possible reasons for this difference may be a higher number of homonyms in OSwedish, combined with a tendency of the composers of the old lexica to distinguish between many contextual nuances within one entry.

4 Tracking lexical change in large-scale material

We focus our case-study on a delimited sets of links, considering only lemmas that appear both in the Old and in the Modern or Contemporary lexica, that is links that were found without using information from the definitions. This intuitively reflects the sets of words that are etymologically related. At the current stage, we further only consider exact matches leaving out lemmas that were linked using spelling variation. While many of the lemmas meeting these criteria are assigned the same set word classes by all lexica, we chose to closer examine the cases when the information about word class differs in one of the resources. In this way we are able to track potential cases of semantic change and grammaticalization. For the rest of this paper, we will take a slightly simplified position in defining and distinguishing between content words and (grammatical) function words, seeing verbs, nouns, adjectives and adverbs (of manner, place and time) as content words, belonging to open word classes, while the other word classes, mainly prepositions, conjunctions, pronouns, sentence adverbs and particles, are considered as functions words (closed class items).

The words that the historical and modern lexica assign to different word classes are mainly interesting as illustrating cases of narrowing and broadening of meaning. The most common case is narrowing: when OSwedish verbs and nouns have several different meanings that are no longer present in the MCSwedish. Examples found in our listing are the verb *rykta* 'manage, nurture, see to, acquire' with the modern counterpart *rykta* 'see to' and the noun *gift* translating to 'marriage' and other meanings related to different gifts, *gåvor*, whose MCSwedish homonyms only means 'marriage' but also the unrelated 'poison'. Cases of semantic generalization also appear to be present in the results of our linking. For instance the noun *flykt* 'flight', only noted as 'bird travel' in OSwedish is far more polysemic in MCSwedish. Another semantic change, namely pejoration of meaning, is seen in the OSwedish noun *tukt* with definitions 'upbringing, courtesy' (positive connotation), and its MCSwedish counterpart meaning 'upbringing,

discipline' (negative connotation).

Table 1 shows the part-of-speech (mis)matches between the lexical resources. To be a candidate in this table, a part-of-speech assigned in one of the lexica have to be lacking in the other. Hence, cases where a lemma is noted as both a verb and a noun in all lexica are not included. Table 1 then shows us the number of overall matches, note however that these not necessarily show a semantic relation. Totally 109 lemma matches are shown in the table. As a first step we

Söderwall	**Dalin/Saldo**	**counts**	**examples**
adj/adv	noun	36	helg, kön, fat
noun	adj/adv	18	arm, hel, skum,
verb	noun	16	glosa, sena
noun	verb	10	bleka, känna,
adj/adv	prep/conj	11	vid, hos, innan, än
adj/adv	verb	7	heta, svara
prep/conj	noun	6	enka, bak, ok
noun	prep/conj	4	mot, und
prep/conj	adj/adv	1	gen

Table 1: Part of speech: Linking matches

may distinguish some patterns. Not all of the links show an (obvious) etymologically relevant connection. For instance, the apparent change from adjective/adverb to noun of some words are rather arbitrary and the high count for this reflect pairs of so called false friends. Examples are OSwedish *helg* (adverb) 'holy' and MC *helg* (noun) 'weekend'. *helg* is a spelling variant to the lemma *helagher* 'holy' in OSwedish. In Modern and Contemporary Swedish the lemma is *helig* and both part-of-speech is thus still productive; the meanings 'holy' and 'weekend' show a clear relation in that weekend correspond to 'holiday(s)'. A similar example is the OSwedish variant *kön* 'proficient' (cf. *kyn* 'sex') and MCSwedish *kön* nn 'sex'. The lemma *kyn* 'sex' is a spelling variant of *kön* in OSwedish and thus ends outside our linking experiment.

Other examples that are not likely to be considered as etymologically related lemmas are the OSwedish preposition *bi* 'by' and conjunction *ok* 'and' and their nominal homonyms in MCswedish; *bi* 'insect' and *ok* 'tool made of wood'. However a great majority of the examples in Table 1 do show a clear relation, mainly referring to contiguity in context (metonyms) or similarity (metaphors).

4.1 Semantic change inside the content word system

The most common matches we find that clearly correspond to relevant cases of semantic change are the transformation from nouns to adjectives and adverbials. Nouns for visible phenomena in OSwedish, such as *blek* 'shine' and *skum* 'twilight' and their adjective homonyms in MCswedish - *blek* 'pale' and *skum* 'dusky' - are clearly related. They correspond to well known word formations from nouns to adjectives (Ljunggren, 1939). Two similar examples that show a more unclear semantic relation are the OSwedish nouns *len* 'hill' and *ny* 'new moon'. Their adjective homonyms in MCSwedish *len* 'soft' and *ny* 'new' are not present in the OSwedish resources. That a meaning as 'new moon' had give rise to the adjective 'new', a kind of generalization, is likely but a closer look at the lemma *len* makes clear that the counterpart *lin* 'soft' is present in OSwedish. Hence the variants *len* 'incline' and MCSwedish *len* 'soft' are hardly related lemmas.

Other interesting semantic correlations are those between nouns and verbs. The Old Swedish

nouns *bleka* 'chalk' and *känna* 'knowledge' should less or more (in that order) be related to the MCSwedish meanings of 'to bleach' (i.e. use chalk, 'kalk' to bleach) and 'to feel' (i.e. some relation between mental and physical states). Interesting changes in the opposite direction are verbs that have nominal counterparts in newer forms of Swedish, such as *glosa* 'interpret' and *sena* 'delay'. The nominal counterpart in MCSwedish *glosa* 'vocable to learn' is clearly related to the action 'interpret'. In the first sight we might imagine a relation even between the verb *sena* 'delay' and the body part *sena* 'sinew' given Dalin's definition, "shiny, round or flat, rubbery strings, through which the muscles are attached". None of the lexical resources discuss this possibility, and a metaphorical relation between delay and rubbery (strings) is of course forced and not likely to be apparent. There are no other matches between verbs and other word classes. It may be due to the fact that the most common change of verbs is the rise of auxiliaries. No such relation is identified, since we do not have enough information on potential auxiliaries in more than one of the resources (Dalin).

Furthermore, we are able to identify some OSwedish adverbs and adjectives with only verbal counterparts in MCSwedish. Exampels are *heta* 'hot', and *svara* 'violent'. However the MCSwedish verbs *heta* 'to name' and *svara* 'to answer' are present in both OSwedish and MCSwedish, which implies that the adjectives in OSwedish are probably an outcome of the more general polysemic nature of OSwedish lemmas, discussed in Section 2. More interesting are the OSwedish adjective as *rena* and *ren*, 'completely', 'honest' which may have given rise to an verbal action *rena* 'to make something clean' (read 'to make something better or more complete, hence more honest'), a possible metaphorical change even though untypical in going from abstract to concrete.

4.2 Cases of grammaticalization

Table 1 also shows us potential examples of grammaticalization. The most striking example of grammaticalization found in our experiment are well known, i.e. the development of a preposition *mot* 'to(wards)' from the OSwedish noun *mot* 'meeting'. The bridging contexts are likely related to the use *om ganga/koma a mot*, 'walk/come to meet', in which the head noun come to be replaced with only the directional meaning. Other examples that might be defined as grammaticalization is the prepositional use of *hos* 'at' and *vid* 'by' in MCSwedish. In OSwedish they are also used as adverbs. Söderwall gives the adverbial meanings for *hos* 'home at' and for *vid* 'at that time, due'. As regards *hos* it is also likely that the noun *hus* 'house' forms the origin of the adverbial use (Hellquist, 1957), although this connection could not be directly identified in our experiment. Even if those words also are used as prepositions in OSwedish, the restricted use to prepositions in MCSwedish indicates a grammaticalization process. Likewise, some adverbs and adjectives have conjunctional counterparts in MCSwedish. Examples are *innan* 'inside, during', *fast* 'fixed' and their conjunctional counterparts *innan* 'until' and *fast* 'though'. They are not noted as conjunctions in Söderwall and we may see this as a grammaticalization process triggered by the conceptual metaphorical change from positions in space to a fixed position in time in the case of *innan*, and possibly from space/time to concessive meaning in the case of *fast* (c.f. Traugott and Dasher, 2002).

The reverse process, called degrammaticalization (discussed and problematized elsewhere, e.g. Andersson, 2008), might be apparent when first looking at the OSwedish preposition *gen* 'against', 'towards', that only has an adjective as counterpart in MCSwedish *gen* 'near'. However, our linking approach has once again obscured the presence of a closer OSwedish counterpart, this time *gin* 'near'. The etymological relationship between the adjective and prepositional

meanings seem to be very fuzzy, (Hellquist, 1957). Other prepositions and conjunctions in OSwedish such as *ok* 'and', *bi* 'at' are discussed above and not likely related to their homonymic nominal counterparts. Another case is the OSwedish preposition *bak* 'behind' that has a nominal homonym in MC Swedish *bak* 'back, spine'. However the latter meaning has the OSwedish counterpart *baker* which with great certainty is the origin with respect to typological and etymological data (Hellquist, 1957). Another interesting example in our linking experiment, OSwedish adjective *enka* 'single, a few' is also noted as a pronoun in Söderwall with the unclear meaning *någon* 'someone'. The MCSwedish counterpart *änka* 'widdow' is likely related to the adjective (ie. 'single') but more doubtful to the pronoun given Söderwall's unsure position of it's pronominal use. Thus, it has to be defined as a change inside the content word system.

5 Conclusion and Outlook

We have introduced a first version of a diachronic lexical resource, connecting three OSwedish lexica to the contemporary resource Saldo, and consequently to the Swedish FrameNet++. To be able to entirely connect the lexica, a considerably higher amount of work would be needed. However, we believe that the results we got by linking up the lexica with a fully automated process will already be of use in analysing historical texts and for processes of change.

Based on the links we found between the historical and modern lexica, we preformed a small study showing some interesting cases of more or less transparent lexical change. Semantic changes inside the content word system such as generalization, specialization, metonyms and metaphors have been identified and further discussed. Potential cases of grammaticalization have also been identified. The most clear cases were the few examples of nouns or adverbs (of space and time) that come to be less or more restricted to grammatical words in MCSwedish. Examples thereof are *mot*, *hos* and *vid*, all of them being used as content words in OSwedish. A few cases of potential degrammaticalization have further been discussed. However, to reach a more solid and deep analysis of the relations identified, it is necessary to complement this study with corpora research.

6 Future work

In Section 3.1 we gave an overview of the tasks at hand for improving the quality and coverage of the link-up of Old Swedish and Contemporary Swedish lemmas. In doing so we may identify several interesting cases of potential lexical change on different levels. For example, a more detailed analysis of the material studied in Section 4 might discover many semantic micro-changes, mainly inside the content word system. Attempts to carry out the linking on a level of senses, as discussed in Section 3, would be highly interesting to evaluate, with tracking conceptual-semantic domains as a possible aim. A related idea is to identify other information in the lexical entries, for example more detailed notes on the use of words, e.g. "obsolete", would be highly interesting to study. Alongside those goals more lexical resources in digitalized form is desirable in linking attempts, such as the already digitalized versions of Swedberg's (Swedberg, 2009) and F.A. Dahlgren's dictionaries (Dahlgren, 1960) on Modern Swedish.

Acknowledgments

This work was carried out in the context of the Swedish FrameNet++ and Diabase at Språkbanken, University of Gothenburg.

References

Adesam, Y., Ahlberg, M., and Bouma, G. (2012). *bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa...* Towards lexical link-up for a corpus of Old Swedish. In *Proceedings of LTHist 2012*.

Andersson, P. (2007). *Modalitet och förändring. En studie av må och kunna i fornsvenska*. Göteborgsstudier i nordisk språkvetenskap 10, Institutionen för svenska språket, University of Gothenburg, Gothenburg.

Andersson, P. (2008). Swedish må and the degrammaticalization debate. In Seoane, E. and López-Couso, M. J., editors, *Theoretical and empirical issues in grammaticalization*, Typological studies in language, pages 15–32, Amsterdam/Philadelphia.

Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., and Kokkinakis, D. (2010). The past meets the present in swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281.

Borin, L. and Forsberg, M. (2011). A diachronic computational lexical resource for 800 years of Swedish. In *Language technology for cultural heritage*, pages 41–61. Springer, Berlin.

Borin, L., Forsberg, M., and Lönngren, L. (2008). SALDO 1.0 (Svenskt associationslexikon version 2). Språkbanken, Göteborgs universitet.

Dahlgren, F. A. (1960). *Glossarium öfver föråldrade eller ovanliga ord och talesätt i svenska språket från och med 1500-talets andra årtionde*. Atelier Elektra, Köpenhamn.

Dalin, A. F. (1853/1855). *Ordbok öfver svenska språket*, volume I–II. Stockholm, Sweden.

de Melo, G. and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

Eriksson, M. (1992). Ett fall av grammatikalisering i modern svenska. Ba i ungdomars talspråk. *FUMS rapport 166*. Institutionen för nordiska språk, Uppsala.

Ernst-Gerlach, A. and Fuhr, N. (2007). Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 333–341. ACM.

Gotscharek, A., Reffle, U., Ringlstetter, C., Schulz, K. U., and Neumann, A. (2011). Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *International journal on document analysis and recognition*, 14(2):159–171.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY - a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590.

Hellquist, E. (1957). *Svensk etymologisk ordbok 1&2*, volume 3. C.W.K Gleerups förlag, Lund.

Hopper, P. and Traugott, E. C. (2003). *Grammaticalization*. Cambridge university press, Cambridge, second edition.

Koolen, M., Adriaans, F., Kamps, J., and De Rijke, M. (2006). A cross-language approach to historic document retrieval. *Advances in Information Retrieval*, pages 407–419.

Ljunggren, K. (1939). *Adjektivering av substantiv i svenskan*. C.W.K Gleerups förlag, Lund.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Pumfrey, Stephen, P. R. and Mariani, J. (2012). Experiments in 17th century English: manual versus automatic conceptual history. *Literary and Linguistic Computing*. Oxford University Press.

Schlyter, C. J. (1887). *Ordbok till Samlingen af Sweriges Gamla Lagar*, volume 13 of *Saml. af Sweriges Gamla Lagar*. Lund, Sweden.

Swedberg, J. (2009). *Swensk ordabok*. Uppsala, Sweden.

Söderwall, K. F. (1884). *Ordbok Öfver svenska medeltids-språket. Supplement*. Lund, Sweden.

Söderwall, K. F. (1953). *Ordbok Öfver svenska medeltids-språket. Supplement*, volume IV–V. Lund, Sweden.

Traugott, E. C. (2001). Legitimate counterexamples to unidirectionality. `stanford.edu/~traugott/papers/Freiburg.Unidirect.pdf`. Paper presented at Freiburg university.

Traugott, E. C. and Dasher, R. B. (2002). *Regularity in semantic change*. Cambridge university press, Cambridge.

Experiments on sentence segmentation in Old Swedish editions

Gerlof BOUMA, Yvonne ADESAM

Språkbanken, Department of Swedish
University of Gothenburg

gerlof.bouma@gu.se, yvonne.adesam@gu.se

ABSTRACT

We present experiments on automatic segmentation of electronic Old Swedish editions into sentence-like units. Our target material is characterized by a great variation in the type of boundaries that are marked orthographically, the extent of boundary marking, and the means of boundary marking. We begin with an exploration of boundary marking in a large, unannotated corpus of Old Swedish texts. Then we show that we are able to improve upon a simple but effective segmenting baseline, using a conditional random field model trained on a manually annotated corpus. A more valuable lesson the modelling teaches us, however, is that we need to address the boundary marking variation explicitly.

KEYWORDS: Sentence-like units, boundary detection, Old Swedish.

Gerlof Bouma and Yvonne Adesam 2013. Experiments on sentence segmentation in Old Swedish editions. *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013.* NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87: 11–26.

1 Introduction

Historical text corpora have recently gained much interest in computational linguistics. Due to differences between the historical and contemporary materials, it is not always the case that we can reuse automatic methods developed for contemporary language on the historical texts. In our efforts to compile and process a corpus of Old Swedish texts (13–16th c.), we have found that tools for modern Swedish cannot be effectively applied. We are facing a situation similar to treating a whole new language. One of the major challenges in this material is the lack of a single orthographic standard. This results, for instance, in a wide variety of spellings – the same word may even be spelled in different ways in the same paragraph. Another consequence is a variety of boundary marking strategies, so that it is hard to determine where one sentence ends and another begins. The methods for marking sentence-like units range from a period followed by a word starting with an uppercase letter, over using slashes, commas, or uppercase letters alone, to no marking at all. Sentence segmentation can be helpful to corpus users exploring the texts and is sometimes necessary for computational reasons.

In this paper, we will give a high-level overview of the different boundary marking strategies in our collection (Section 2) and present machine learning experiments on the sentence boundary detection task in Old Swedish on a hand-annotated corpus (Section 3). In the remainder of the current section, we will motivate our interest in the sentence boundary detection task and discuss some relevant previous work.

1.1 Motivation

Segmentation of a text into tokens and sentences is the first step in the traditional NLP pipeline model, minimally consisting of segmentation followed by part-of-speech tagging followed by parsing. The interesting and hard problems occur further down in the pipeline, whereas the problem-free execution of the first step is typically taken for granted. Irrespective of whether sentence segmentation really is a 'solved problem' in modern, professionally written and edited English, it is certainly not the case for our Old Swedish material. Currently no segmenter exists and, as we hope to make clear in the rest of the paper, no single strategy exists that could give good enough results to feed the NLP pipeline. As many annotation tools at later stages rely on having bounded strings to operate on, the pipeline model breaks down without a way to segment the data early on.

Given the lack of clear boundary marking strategies in the material, we might consider deviating from the traditional NLP pipeline model. If we have access to, for instance, a parser that can handle unbounded strings, we could use this to annotate the stream of tokens directly and skip the lower-level segmenting task. If a graphic sentence-like segmentation of the text is desired for other reasons, one could then reconstruct this from one of the other levels of annotation. For instance, the maximal syntactic unit, the *macrosyntagm* (Loman and Jørgensen, 1971), could serve as a suitable unit.

Although we certainly hope to investigate this option in the future, the investigation and experiments presented in this paper are still valuable for a number of reasons. Even if we would have the type of part-of-speech taggers and/or parsers available that do not need pre-segmented data, these tools will probably benefit from information about boundary marking present in the input. This information could come from a stand-alone model, whose predictions are taken into account by the higher level annotation tool. Alternatively, the higher level tool will need to learn about boundary marking itself, in which case we need to supply it with knowledge about

the possibilities and their distribution. In either case, the research presented in this paper will be of help.

Finally, there is also a non-computational reason for the work in this paper. Segmentation is very helpful for corpus users. Searching an unsegmented text and inspecting hits in an unbounded string is cognitively straining. Put concretely, one doesn't know where to start reading. Higher level tools, like a parser, that could give us a useful segmentation as a byproduct have not yet been developed for Old Swedish, nor do we have the necessary annotated data available. The data and dedicated segmentation models investigated in this paper could immediately help to improve the presentational situation.

1.2 Background

Following Stevenson and Gaizauskas (2000), we distinguish between punctuation disambiguation and sentence boundary detection. Segmenting a text into sentences or sentence-like units sometimes reduces to the much more restricted and arguably easier task of punctuation disambiguation. The task is then, given a number of target delimiters like full stop, exclamation mark, question mark, to decide whether they end a sentence or not. In the case of the full stop in English orthography, this means for instance choosing between a sentence final full stop and an abbreviation final period. This restricted task can be solved with near perfection for well-behaved corpora. Reported error rates are well below .5% on professionally written data (Mikheev, 2002; Gillick, 2009). Even completely unsupervised systems are able to achieve error rates below 2% (Kiss and Strunk, 2006) in this task on the same data (see Gillick, 2009, for a comparison). The idea that punctuation disambiguation error rates give a good idea of the performance of the reported systems as segmenters has been challenged, though. See Read et al. (2012) for discussion and a more realistic assessment of the state-of-the-art.

Sentence boundary detection is a more general and harder task in which a language stream (text, transcribed speech) has to be segmented into sentence-like units on the basis of a mixture of information sources, without being able to reduce the segmentation into the recognition or disambiguation of a single signal. This problem description applies in the case of speech data, where the task is often formulated as a labelling problem. Each token in the speech stream is either a boundary token or a non-boundary token. The models for this decision use different types of information, like n-gram language model probabilities and prosodic features taken from the acoustic signal (Gotoh and Renals, 2000; Liu et al., 2005). The latter report a token-based error rate of just below 3.5% for detecting boundary tokens in broadcast news data.[1]

For our Old Swedish material we are dealing with the sentence boundary detection task, because writers mark a variety of different types of boundaries besides sentence-like units, with a variety of marking strategies. In a sense, punctuation symbols and capitalization are to us what prosodic features are to the speech segmentation researcher: an important but uncertain source of information to supplement the lexical level. There may be overall tendencies in the data, but the observed variation makes these non-obvious. Variation is seen looking across documents – which span a time frame of about 3 centuries, are produced at many different places and, not unimportantly, have passed through the hands of several editors – but also within them. Consider the following four sentences. For the larger part, the writer of the document uses a

[1] Note that there is no way of directly comparing the error rates between the two tasks, although it might be possible to convert the punctuation disambiguation error rate by taking the rate of occurrence of disambiguation points into account. Also recall that error rates need to be seen in relation to their majority class baselines.

slash '/' to mark (presumably) some kind of prosodic phrasal break. However, in the last part of the document, no marking whatsoever is used. The sentences below are taken from the part in the document where this strategy switch occurrs. (Henceforth, '‖' marks sentence boundaries inserted as annotation after the fact.)

(1) sua ma han þem æpte sinum vilia bøgha / at þe mago kallas oc vara / þe hælgo kirkiu dorakroka / ‖ mz þe hælgho kirkio gulfe / menas biskopa / oc værulsleke klærka / ‖ þera giri ær sua diup / þæt þær ma ingte i grynna / oc þera høghfærfe oc skøro lifærne gar af fragh [...] ‖ þætta ma pafin an han vil mykyt at bættra ‖ raþe allum biskopum þy følghia i gozs oc allum andrum þingum sum þu hørþe at honum var raþet siælfum at gørra
'Thus he may bend them after his will, so that they may [truly] be called the holy church's door hinges. With the holy church's floor, the bishops are meant and secular clergy. Their greed is so deep that nothing can reach its bottom, and their vanity and sinfulness is infamous [...]. This the pope can make a lot better, if he wants. Advise all bishops to do in questions of property and all other things like you heard he himself was advised to do.' (*Brigitta-autograferna*).

Other prominent marking strategies involve full stops, commas, colons, semi-colons, capitalization and combinations of a delimiter and capitalization. Even if our task reduces to punctuation disambiguation for a given part of a given document, this is by no means the case in general.

To our knowledge, historical sentence boundary detection is not a well researched area. Previous work has followed the speech processing literature by treating the problem as a labelling task. (Huang et al., 2010) segment Classical Chinese and 19th century Chinese, material which lacks any marking so that all information has to come from the lexical text level. They report f-scores of around .84 for their data set. (Petran, 2012) reports experiments on an artificial historical data set, created by removing capitalization and punctuation marks from a modern German newspaper corpus. Using part-of-speech information, he is able to reconstruct sentence breaks with an f-score of .65 and without part-of-speech tags with an f-score of .50. He also shows that reconstructing smaller units like (syntactic) clause and NP/PP chunk is an easier task which may be worthwhile to use as an intermediate step.

We will follow this research and treat segmentation like a labelling task. Before we present our experiments, we will give an overview of different boundary marking strategies in our unannotated corpus.

2 Exploring Sentence Boundaries at Text Level

Our unannotated data – the corpus we eventually wish to make available with annotations – has been supplied by Fornsvenska Textbanken[2] and consists of close to 3 million words in about 150 texts. The corpus contains fiction, legal and religious prose, with texts from the 13th to 16th century and ranging from 200 to 200 000 tokens in length. We have excluded verse from our investigations, because it comes with its own set of conventions.

To get an overview of various boundary marking clues in the texts, we start by extracting non-alphabetic characters, such as *: ; , /* . · – and ¶. We then calculate token-delimiter ratios for each text, that is, the frequency of occurrences of a single delimiter, expressed as the average

[2] `http://project2.sol.lu.se/fornsvenska/`

Marker:	.	.+Cap	,	,+Cap	/	/+Cap	Cap
T-d ratio:	5.22	14.20	86.74	86.90	86.90	86.90	12.82

Nv ærum wi skyldughi at tiunda af alle sæþ ware. || Tiunþ scal i akrum af tæliæs || byriæ at þem skyli fyrstum vp skars oc af bars oc tæliæ swa sum til tiu. || byriæ owan at akri oc lyctæ wiþ ændæ. || byriæ ater þær wiþ ændæ tæliæ. oc lyctæ owan warþa. || fari swa æ mæþæn akra winnæs. || ei scal korn mællum akra bæræ. || Tiund þæsse scal i þry skiptes || taki prester en skyl. || twa före bonde hem til sijn. oc i þry skipti. en lot kirkiuni annæn biscupi. þriþiæ fatökum mannum. || bönder aghu presti til tiunþ sinnæ sighiæ þa þe laþa wilia vm þrea sunnudagha oc lagha wærn vm halda. || Warþar hon ei gömþ innæn þe þriæ sunnudagha. oc warþer hon ætin eller spilt. böte þen ater tiunþ sum þet vlti at prester scaþa fik. || Hawi oc prester siælwer scaþa æn han ei gömer innæn þe þriæ sunnudagha. oc ærum wi skyldughir at tiunda af alle þe sæþ sum man arwþer i iorþena.

Figure 1: Token-delimiter ratios for *Södermannalagen* with an example fragment.

number of tokens between them. We also consider capitalization as a boundary marking strategy, and calculate corresponding token-delimiter ratios for capitalization alone and combinations of any of the punctuation symbols and a following capital.

We hope the token-delimiter ratios can give us clues about the type of segments they delimit. We use this with caution, however, since a single text may use more than one kind of delimiter or not mark all segment boundaries; in both cases, segments are actually shorter than the token-delimiter ratio indicates.

Figure 1 shows an example text and token-delimiter ratios for the most common markers. In this text, capital and period followed by a capital have token-delimiter ratios in the 10–30 range, a range we consider to hint at delimiters that are used to mark sentence-like units. Looking just at the surface features of the example text, we see that in this fragment, capitals reliably indicate sentence boundaries. Periods alone, however, mark a much smaller unit. This is reflected in the low token-delimiter ratio for periods.

Of the 149 texts, about three quarters contain one or several delimiters with a token-delimiter ratio in the 10–30 range. From this way of looking at the data, capitalization appears to be the most prominent single clue, falling in the 10–30 range for 94 documents and rarely exceeding 50. In the machine learning experiments of the next section, we will also see that capitalization is an important clue, although it is neither universal nor perfectly reliable.

In Figure 2 is a fragment from a text where none of the studies delimiters have a token-delimiter ratio below 70. The most frequent markers are comma and capitalization, although the high token-delimiter ratios suggest that many of the sentence boundaries will simply lack surface marking. This is illustrated in the fragment in the same figure. Although some of the sentences end in a comma, most of them have no boundary marker. This clearly demonstrates that segmentation for this material does not reduce to punctuation disambiguation.

If we order the texts after their approximate production period, we notice an interesting trend. Texts where commas have token-delimiter ratios in the 10–30 range are more common in the later periods. This fits nicely with claims made in Svensson (1974), who names the comma as the most common delimiter – marking 'longer pauses' – at the end of the Old Swedish period. Svensson also characterizes the slash '/' as a marker for pauses. We do note, however, that 16

Marker:	.	.+Cap	,	,+Cap	/	/+Cap	Cap
T-d ratio:	435.80	435.80	72.63	326.85	435.80	435.80	77.82

Thaa kesarinnan fik höra ath then wnga herrän war kommen thaa tilreddhe hon sigh mz iomfrum och klädom som hon aldra bäst kunne och kom gaangande til konungen och tilhans son ther the saatho baaden til samman || konungen bad hänne sätia sig när sonnen || kesarinnan sade til konungen herra är tättha edher son som saa länghe hafuer borto waridh när the wisa mästara || konungen sade ya män jak kan ekke wettha hurw thz gaar til thy han wil inthe tala, || thaa sadhe hon herra antuarden honnom mik jak skal wil wäl göra honnom talande och togh honom widh haandena och wille hafuan mz sigh, || thaa warde han sig och wille ekke mz, || fadren bad honom gaa mz hänne || thaa negh han sinom fader ödhmyuklighan och war honom lydoger || kesarinnan ledde honom in j en kammara och badh alla wtgaa och satte honnom oppaa en sänga stok när sigh och sadhe, hiärtans käre diocleciane, huad stor aastundan och länktan jak hafuer äpter tik haft, fran then första dagh

Figure 2: Token-delimiter ratios for *Sju vise mästare C*, with an example fragment.

texts have a token-delimiter ratio for '/' below 30, of which only 2 are below 10. Its alleged status as a pause marker is therefore not obvious from the data, as we would expect to see a larger proportion of low token-delimiter ratios.

Finally, larger units of texts are often marked with the paragraph sign '¶'. It occurs in more than half of the texts. In half of these cases, the token-delimiter ratio falls between 50 and 500, and in a quarter the token-delimiter ratio is above 1000. The wide range of ratios suggests that the paragraph sign may delimit larger units of different types.

A special way of marking the end of a larger unit is found in *Peder Månsson's Bondakonst*, which uses the combination ':-'.

(2) Ta oxana haffwa draghit oc lösas gnides oc strykes wäl halsen oc ryggen theras mädh handommen, oc wpdrage alla stadz wäl skinneth fran ryggenom || ey lantandis thet lodha widher ryggen thy then sywkkdommen är them ganskans ondher som kallas hwdbända || oc ärw the mykith hethe tha latis ey til ath ätha för än the wenda ather ath flämtha oc swetten är bortagangin, || oc giffwes them litith j sänder äta oc swa meer oc meer, || sidan gifwes them drikka, oc swa giffwes them nogh ätha oc rökthes wäll:-
(*Peder Månssons Bondakonst*)

3 Segmentation as tagging

In addition to the unannotated corpus used in the previous section, we also have access to a smaller, hand-annotated corpus, which we use to train a statistical segmentation model. We treat the sentence boundary detection task as a sequence labelling task, assigning to each token a label that indicates whether it starts a sentence or not.[3]

[3]As pointed out by an anonymous reviewer, one can also project sentence boundaries for parallel data such as bible texts, to supplement the manually annotated training data (see, e.g., Haug et al., 2009). This should be explored in future work.

Following (Huang et al., 2010; Petran, 2012) we distinguish five tags in our labelling task. These five tags are:

S: 'singleton', one-token sentence;
L_0, L_1: 'left', first and second token in a sentence, respectively;
M: 'mid', tokens in the middle of the sentence;
R: 'right', last token in a sentence.

Each sentence will be labelled with a tag sequence conforming to $S|L_0(L_1M^*)?R$. Compared to a minimal tag set, with only the distinction sentence beginning vs everything else, the extra state information improves segmenting results (Huang et al., 2010).[4] When applying the resulting labellers to the segmenting task, a new segment should be started for each occurrence of S and of L_0.

3.1 Data preparation

We construct training data from the HaCOSSA corpus (Höder, 2011), a corpus of Old Swedish texts with partial syntactic annotation, consisting of 13 documents ranging between 500 and 32k tokens. HaCOSSA contains among other things annotation of main/subordinate clauses, which we use to define sentence-like segments. Each left edge of a main clause starts a new segment, as do ¶ marks and text structure elements such as titles and paragraphs. In HaCOSSA, not all subordinate clauses are attached to a main clause. Our conversion of clause annotation to sentence-like structuring in effect integrates all unattached subordinate clause into the first reachable main clause to their left. The resulting corpus has just over 8k sentences with about 137k tokens including punctuation (16.5 tokens/sentence).

The documents in the corpus come from different times, different writing schools and different editions, which means that there is considerable variation in orthography. Variation is also found within documents. The wealth of boundary marking strategies is of course the main focus of this paper, but the variation in spelling is an unwanted source of trouble in our machine learning experiments. Its effect is reduced generalizability. First and foremost, spelling variation will lead to data sparseness. Even if a certain word is a good clue for sentence boundary marking throughout the corpus, it might be that the machine learning fails to pick up on this, because the word is spelled in many ways. Secondly, if the model associates certain spellings of a word with boundary marking, this may create problems for evaluation through cross-validation. Using spelling as an intermediary, the model effectively learns from which document and boundary marking convention a sequence comes, which leads to cross-validation giving an inflated idea of the models performance on new data.

As a step towards addressing these problems, we apply a very crude spelling normalization to the texts. The normalization uses simple character mappings, derived from corpus inspection. They are listed in Table 1. The mappings reduce the type count from just under 21k to just over 16k. For future work, we intend to investigate the application of more refined spelling variation handling techniques (Adesam et al., 2012) to the training material.

[4] Even though it is possible that the CRF tagger assigns a nonsensical tag sequence, such as R L_1, this seldom occurs in our experiments. Petran (2012) makes the same observation. Moreover, nonsensical tag sequences do not affect our ability to segment text on the L_0 tags. Technically, it should be possible to restrict the CRF tagger to only yield legal sequences, but this was not explored in the experiments for this paper.

Raw	Norm	Raw	Norm	Raw	Norm
aa	a	gh	g	yy	y
dh, þ	d	ii, ij, j, jj	i	ø, øø, öö	ö
c, ch, q	k	oo	o	æ, ææ, ää	ä
ee	e	th	t	åå	å
ff	f	u, uu, vv, w, ww	v		

Table 1: Spelling simplification mappings used in the normalized data experiments.

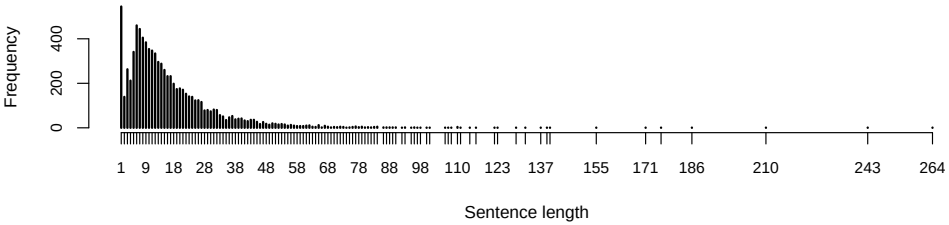

Figure 3: Histogram of sentence lengths in the annotated corpus

3.2 A first look at the data

Before we go into the experiments themselves, let us take a closer look at the data. A histogram of sentence lengths is given in Figure 3. As mentioned above, the mean lies at 16–17 tokens per sentence, and as can be seen from the histogram, the modal sentence has 6–7 tokens. The peak at sentence length 1 consists mainly of ¶-marks, which always constitute their own segment. These are thus trivial to annotate with the S tag.[5] The slight peak at 3 is due to segments of the form ·*XLII*·,[6] which appear in just one of the documents in the set. We note however, that the notation ·NUMERAL· and ·*i*· (either roman numeral 1 or the preposition 'in') is common even sentence internally. Although we can expect a right-tailed distribution of sentence lengths, the extremely long sentences are likely to be an artifact of the way we handled unattached subordinate clauses.

Table 2 gives an overview of the association between lexical items and the L_0, L_1 and R tags, both in terms of the conditional probability of the tag given a token and of the token given the tag. The top ten item lists for each tag were created by sorting the tokens after the product of these two conditional probabilities.[7]

As we can see in the table, the start of a sentence (L_0) is associated with a mix of discourse particles and (predominantly personal) pronouns. The particle *ok* 'and/too' is both very frequent as a first token in a sentence and fairly reliable as a beginning of sentence clue. A more reliable

[5] In the evaluation in the next section, we only look at the accuracy of recognizing the start of a new multi-token sentence L_0. The fact that ¶ is easily recognizable as an S and the preceding token as an R does therefore not skew the evaluation results compared to incorporating ¶-marks into a previous or following segment.

[6] Depending on the text and edition, these dots may also rest on the baseline like (modern) full stops. We show the centred dot examples to set them apart from modern punctuation in the body text.

[7] This product itself is not shown in the tables, as it is just a means to select examples and not a quantity of interest in this discussion. Sorting tokens by the product of p(Tag|Token) and p(Token|Tag) gives the same ordering as the association measures PMI^2 and *geometric mean*, known from the collocation extraction literature (Evert, 2005).

w	$\sum w$	$p(L_0\|w)$	$p(w\|L_0)$	w	$\sum w$	$p(L_1\|w)$	$p(w\|L_1)$	w	$\sum w$	$p(R\|w)$	$p(w\|R)$
ok 'and/too'	9002	.24	.28	*är* 'is'	1516	.22	.04	/	4341	.44	.25
nv 'now'	603	.63	.05	*skal* 'shall'	529	.26	.02	.	4647	.35	.21
ta 'then'	1023	.28	.04	*skalt* 'shall'	153	.48	.01	*svarade* 'replied'	134	.51	.01
sidan 'since'	286	.45	.02	*talar* 'speaks'	125	.42	.01	*sigiande* 'saying'	59	.56	.00
o 'o' (VOC)	132	.58	.01	*äpter* 'after'	183	.33	.01	*etketera* 'etc.'	13	1.0	.00
iak 'I'	967	.20	.02	*sagde* 'said'	272	.27	.01	*ketera* '[et]c.'	13	.92	.00
ty 'for/it'	1171	.15	.02	*vil* 'wants'	163	.33	.01	*sagde* 'said'	272	.18	.01
ter 'there'	663	.19	.02	*ty* 'for/it'	1171	.12	.02	,	307	.17	.01
han 'he/him'	2354	.10	.03	*lifde* 'lived'	55	.53	.00	*sigiandis* 'saying'	15	.73	.00
tz 'it'	820	.16	.02	*man* 'person'	332	.21	.01	*döttir* 'daughter'	17	.65	.00

Note: For reasons of space, morphological features are not given explicitly and only approximated in the English translations. Spelling has been normalized according to the rules in Table 1.

Table 2: Tag-token associations for the L_0, L_1, and R tags in the annotated corpus.

clue, but one that is genre restricted,[8] is the word *nv* 'now', which is used to start a new case in a legal text. A common continuation after such a case starting sentence, is a sentence that starts with *ta* 'then', which introduces the consequences and rules applicable in the case. This combination is illustrated in the following example:

(3) **Nu** hittir man fynd innæn allmannæ leþ. hwat fynd þæt hælzt ær. || **þa** aghi þær aff hwarn attundæ pænning.
'Now someone finds something on a public road, whatever the found thing is. Then [they] have a right to an eighth of its value.' (*Upplandslagen*)

Looking at tokens that associate with L_1, we see mainly finite verbs, a reflex of the verb-second tendency in Old Swedish. The presence of *äpter* 'after' shows a limitation of defining tokens as graphical words, as almost all these cases are part of the discourse connective *ter äpter* 'thereafter'.

In the rightmost table, tokens that hint at the end of a sentence (R), are either punctuation marks or *verba dicendi*. The latter is a result from the HaCOSSA annotation guidelines, where direct speech is seperated from the speech verb (Höder, 2011, s3.6). Note that *etketera* 'etcetera' and its graphically split variant *et ketera* almost exclusively appear at the end of a sentence.

In the previous section, token-delimiter ratios suggested capitalization is a promising candidate for a general segmentation clue. Indeed, in the annotated dataset, capitalization is strongly associated with the L_0 tag. Of all capitalized words, 67% start a new sentence and, vice versa, of all sentences, 57% start with a capital.

3.3 Machine learning experiments

The data inspection of the previous subsection gives us an idea of the kind of features that could be useful when training a statistical labeller. It also gives us an impression of the kind of performance that we should minimally expect from a system. We can read the conditional

[8] The only text of this kind in the HaCOSSA corpus is an edition of *Upplandslagen*. However, the same pattern can be observed in some of the other old laws, too.

Feature	Description
1	current token lower case string
2	Is the current token capitalized?
3	bag of two preceding tokens lower case strings
4	following token lower case string
5	Is the following token capitalized?
6	two character suffix of the current token
7	current token category (numberlike, punctuationlike, wordlike)
1×2	current token × is the current token capitalized?
2×3	preceding token × is the current token capitalized?
3×7	preceding token × current token category
4×7	following token × current token category
1×2×3	current token string × preceding token string × is the current token capitalized?
M	preceding × current tag (first-order Markov assumption)

Table 3: Feature descriptions for the {L_0,L_1,M,R,S} labelling task.

probabilities in the tables and text above as precision and recall. So, a classifier that would assign L_0 to all and only occurrences of the token *ok* would have a precision of .24 and a recall of .28. More interestingly, the classifier that tags all and only capitalized words as L_0 would have a precision of .675 and a recall of .573, giving an f-score of .620. This latter classifier will serve as a baseline in the experiments below.

The association between tokens and certain tags strongly suggests that lexical information should be included as a feature in our tagger. The lexical level is also the only information source that can help us tackle the examples of sentences that lack any surface marking, for which we saw examples in the previous section. Capitalization information is also included as a separate feature. The observation that L_1 associates with finite verbs is hard to model in the absence of some kind of part-of-speech or morphological tagging. However, as a very rough approximation, we include token suffixes into the model: the last 2 characters of each token are used as a feature. In our discussion of sentence lengths, we noted that numerals in the text were often accompanied by special punctuation. To help the model recognize these cases, we include a category feature that divides tokens into one of three categories: *numberlike* (roman numerals or indistinguishable from such), *punctuationlike* (all non-alphabetic characters) and *wordlike* (everything else).

Looking at the results above and at the previous section, it is clear that the model should also be able to use combined strategies, such as: assign R to every punctuation mark that is followed by a capitalized word. To some extent, such tendencies may be captured by making a tag's likelihood depend on the previous tag, but more fine-grained information is possible using conjoined features.

We trained a linear conditional random field tagger with the features listed in Table 3.[9] We evaluate the tagger using cross-validation with two different regimes. First, we perform random

[9]To construct the tagger we used the CRF++ package (`http://chasen.org/~taku/software/CRF++`), with smoothing factor c=0.1 and a feature count threshold f=3.

		Random 10-fold crossval			Leave 1 text out		
Data	Model	P	R	F_1	P	R	F_1
normalized	all features	82.9	66.4	73.7	76.0	58.2	65.9
	no capitalization	79.0	52.0	62.7	71.6	39.3	50.7
	no categories	82.7	66.0	73.4	76.1	58.3	66.0
	no context	77.6	57.7	66.2	73.4	52.7	61.4
	0th order	80.8	62.2	70.3	75.7	55.4	64.0
raw	all features	82.2	65.1	72.7	70.6	58.3	63.9
smurfed	all features	71.2	57.2	63.4	70.3	55.9	62.3
baseline (capitalization)					67.5	57.3	62.0

Table 4: Sentence segmenting model comparison (micro-averaged L_0 tag precision, recall and f-score, per evaluation regime).

tenfold cross-validation. Each paragraph is randomly assigned to a fold, giving 10 more or less equally sized folds. However, since the boundary marking strategies are likely to differ strongly from text to text, a leave-one-document-out evaluation regime will give a more realistic estimate of our segmenter's performance. Note, however, that since document sizes span 2 orders of magnitude, the resulting 13 folds are very uneven in their training/testing data set size ratios.

By comparing results between regimes, we can draw conclusions about how general the models are. If adding or removing features affects the results in the same way in both regimes, this means that the model is not sensitive to properties particular to one document, that is, the model generalizes well over document types. Conversely, if we observe an effect in one regime but not the other, this difference points to a lack of generalizability of the features.

Table 4 shows a comparison of models with access to different kinds of information under the two evaluation regimes. Looking first at the difference between the two regimes, we conclude that random cross-validation gives considerably higher averages than document-based cross-validation – a clear sign that overall the models are struggling to find good generalizations that hold across texts. The spelling normalization, rudimentary as it may be, does help in this respect. If we compare the all-features model trained on original (raw) text to the one trained on normalized text, we see an improvement of more than 5 points in precision in the per-document evaluation regime for the latter.

Focussing on the normalized data and the document-based validation,[10] we see that the capitalization information is the most informative feature, followed by the context information, that is, information about preceding and following tokens. The categorization information does not seem to add anything to the model that can be generalized across documents. The all-features model outperforms the capitalization baseline with almost 4 points in f-score, mostly due to a greatly increased precision.

To see how much the models benefit from the lexical level, that is, the words themselves and not

[10]Note that although the averages in the random cross-validation are much higher in this group of experiments, the trends are very similar.

				Precision		Recall		F_1-score	
	$\sum word$	$\sum sent$	$\sum par$	Bl	Af	Bl	Af	Bl	Af
1	1 074	55	1	-	▪42.9	0.0	▪ 5.5	-	▪ 9.7
2	20 241	1 083	67	68.4	▪84.7	64.9▪	54.3	66.6▪	66.2
3	14 482	631	50	55.5	▪78.5	63.5	▪67.0	59.3	▪72.3
4	3 488	107	19	47.7	▪75.7	76.6▪	72.9	58.8	▪74.3
5	563	38	2	75.0	▪87.2	63.2▪	34.2	68.6▪	49.1
6	1 447	115	14	76.6	▪87.2	42.6▪	29.6	54.7▪	44.2
7	7 439	385	39	48.7	▪68.7	82.6	▪83.4	61.3	▪75.4
8	35 288	2 069	52	91.6▪	86.9	45.6	▪53.7	60.9	▪66.4
9	9 972	345	24	60.1	▪77.2	60.3▪	59.7	60.2	▪67.3
10	22 376	1 566	74	73.8	▪75.0	77.7▪	72.7	75.7▪	73.8
11	18 497	1 229	291	79.6▪	58.7	32.0	▪42.4	45.6	▪49.2
12	3 223	114	11	37.7	▪65.7	78.1▪	57.0	50.9	▪61.0
13	3 255	64	11	22.8	▪63.5	71.9▪	62.5	34.6	▪63.0
			Micro average	67.5	▪76.0	57.3	▪58.2	62.0	▪65.9
			Macro average	61.5	▪73.2	58.4▪	53.5	59.8	▪61.8

Table 5: Per document comparison of the baseline (Bl) and the model including all features (Af). The top results per document are highlighted.

just surface clues like punctuation and capitalization, we also trained models on delexicalized (or: 'smurfed') data. All non-punctuation tokens are replaced by one and the same word, but capitalization is kept as in the original. The loss in both precision and recall shows that the models can successfully use and generalize lexical information. However, the fact that the two evaluation regimes now lead to much more similar results than in any of the other data/model combinations suggests that, even though the lexical information is potentially very useful, it is also particularly hard to generalize.

In terms of error rate and NIST score,[11] the all-features model in the leave-one-document-out regime has an error rate of .034 and a NIST score of .602. The capitalization baseline has .040 and .703 respectively. Labelling none of the tokens as L_0 gives an error rate of .057.

The advantages of using the statistical model over the simple baseline may seem modest, given the amount of information that goes into them. However, even though the capitalization baseline gives a good overall performance, it may be that there are large differences on a per-document basis. By combining different strategies, the statistical model could in principle guard us against the variability between documents. Table 5 breaks down performance per fold in the leave-one-document-out evaluation regime, and gives macro averages in addition to the micro averages used thus far.[12] The table also gives the size of the left out documents, to give an idea of the influence of training set size on performance – when the left out document is up

[11] On the familiar true/false positive/negative contingency table, these are defined as follows: Error rate is $(\textit{false positives} + \textit{false negatives})/N$; NIST score, $(\textit{false positives} + \textit{false negatives})/(\textit{true positives} + \textit{false negatives})$ (Liu and Shriberg, 2007). Note that the NIST score is 1.0 if we label everything as negative and may be above 1.0.

[12] Macro averages: an unweighted average over the precision and recall scores of the folds. Micro average: an average where each fold is weighted by the number of tokens in the test document.

to a fourth of the total data size, one might expect to notice an effect of the reduced training set size. However, no such correlation is visible in the table, suggesting that even if such effects may exist, they are dwarfed by the between-document variability.

The baseline has better macro-averaged recall, but in the other averages, the statistical model outperforms the baseline. Looking at the folds, we can see that the statistical model generally is more precise, whereas the baseline has a higher recall in over half of the folds. Although the model has learned to include other clues, it has also become more conservative in labelling tokens as L_o. The better micro-averaged recall of the statistical model is due to its better recall on some of the larger folds, like 3, 8 and 11.

The advantage of having other clues does show in fold 1, which does not use any capitals. The text uses '/' to mark a smaller unit (token-delimiter ratio: 8.4).

(4) enne persona syntis vakande / oc eg sofande / sum hon vare i eno palacio / || oc i fræste væginne syntis en sol / myok stor || framan fore solinne / varo satte sua sum tuu predikaro stola / annar høghra vaghin i þe palacio / oc annar a vinstra væghin /
'One person appeared awake, and not sleeping, as if she was in a palace. And in the furthest wall a large sun could be seen. Before the sun were set two pulpits, one against the right wall of the palace, the other against the left.' (*Birgitta-autograferna*)

Whereas the baseline is obviously useless in this case, the model has picked up enough to correctly identify a couple of L_0s in the material – a recall of just over 5% in this case corresponds to only 3 sentences. Of course, a model trained on the other documents, which *do* provide capitalization information, is likely to give too little weight to cases where the alternative strategies are used alone. This suggests that a model that cannot use capitalization as a feature should do better on a text like fold 1. Indeed, the 'no capitalization' model that fares poorly overall (Table 4), correctly identifies 10 sentence boundaries, a recall of almost 20%, without any loss in precision on this fold.

Although the text in fold 1 is too small to draw any hard conclusions, it does suggest a strategy for coping with the between document variation: using statistics over the text like the token-delimiter ratio for different markers, we can try to select a feature set that is likely to fit the data well. Working out the details of such a strategy and evaluating it will have to remain the subject of future work.

4 Conclusion and Outlook

The segmentation of historical texts into sentence-like units is a useful but hard task. In this paper, we have given a brief overview of the kind of boundary marking strategies we observe in our corpus of Old Swedish. Furthermore, we showed that a model that combines clues from punctuation, capitalization and lexical content is able to improve upon a simple capitalization baseline, especially in terms of precision.

As it stands, the segmentation quality of the models described in this paper is lacking. We can distinguish three use scenarios: for the presentation of corpus query results to users, as a first step in an automatic processing pipeline, and as a preprocessing step in a manual annotation task. In the latter scenario, the segmentation can be adjusted by the annotator and we judge the segmentation quality to be high enough to be helpful. In the first two cases, however, quality is crucial and currently too low.

The main finding of the paper is that the variation between documents is a real obstacle towards better performance. This variation comes in two flavours. First, we have variation in boundary marking strategies. In future work, we hope to be able to use statistics over the occurrence of known boundary markers in a text to choose a model that is likely to give good results for that particular style of marking. Secondly, there is variation in the spelling of words, which means that the models have a hard time picking up lexical clues. It is remarkable, however, that even the simple character mappings used in our experiments make a noticeable difference. In future work, we will investigate more careful and principled spelling normalization to address this problem better.

In the presentation use case mentioned above, we could sacrifice some recall for precision. At a recall of .5, segments will be on average twice as long as they should be, but this is still better than presenting an unsegmented text to the user. Trading recall for precision is generally possible in statistical models, and we will look at methods for boosting precision in future work.

Acknowledgments

We thank three anonymous reviewers for their comments. 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

Adesam, Y., Ahlberg, M., and Bouma, G. (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.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, IMS Stuttgart.

Gillick, D. (2009). Sentence boundary detection and the problem with the U.S. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 241–244, Boulder, Colorado. Association for Computational Linguistics.

Gotoh, Y. and Renals, S. (2000). Sentence boundary detection in broadcast speech transcripts. In *ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium*, pages 228–235, Paris, France.

Haug, D. T. T., Jøhndal, M., Eckhoff, H. M., Welo, E., Hertzenberg, M. J. B., and Müth, A. (2009). Computational and linguistic issues in designing a syntactically annotated parallel corpus of indo-european languages. *Traitement Automatique des Langues*, 50.

Höder, S. (2011). *Phrases and Clauses Tagging Manual for syntactic analyses of Old Nordic texts encoded as Menotic XML documents (PaCMan)*. University of Hamburg, Hamburg. Version 2.0.

Huang, H.-H., Sun, C.-T., and Chen, H.-H. (2010). Classical Chinese sentence segmentation. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 15–23.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Liu, Y. and Shriberg, E. (2007). Comparing evaluation metrics for sentence boundary detection. In *ICASSP*.

Liu, Y., Stolcke, A., Shriberg, E., and Harper, M. (2005). Using Conditional Random Fields for sentence boundary detection in speech. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 451–458, Ann Arbor, Michigan. Association for Computational Linguistics.

Loman, B. and Jørgensen, N. (1971). *Manual for analys och beskrivning av makrosyntagmer*. Studentlitteratur, Lund.

Mikheev, A. (2002). Periods, capitalized words, etc. *Computational Linguistics*, 28(3):289–318.

Petran, F. (2012). Studies for segmentation of historical texts: Sentences or chunks? In Mambrini, F., Passarotti, M., and Sporleder, C., editors, *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities*, pages 75–86, Lisbon.

Read, J., Dridan, R., Oepen, S., and Solberg, L. J. (2012). Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India. The COLING 2012 Organizing Committee.

Stevenson, M. and Gaizauskas, R. (2000). Experiments on sentence boundary detection. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 84–89, Seattle, Washington, USA. Association for Computational Linguistics.

Svensson, L. (1974). *Nordisk Paleografi*. Number 28 in Lunda studier i nordisk språkvetenskap, serie A. Studentlitteratur, Lund.

The *Anselm Corpus*: Methods and Perspectives of a Parallel Aligned Corpus

Stefanie Dipper[1], *Simone Schultz-Balluff*[2]

(1) Department of Linguistics, Ruhr University Bochum
(2) German Department, Ruhr University Bochum

dipper@linguistics.rub.de, simone.schultz-balluff@rub.de

ABSTRACT

This paper presents ongoing work in the *Anselm project* at Ruhr-University Bochum, which deals with a parallel corpus of historical language data. We first present our corpus, which consists of about 50 versions of the medieval text *Interrogatio Sancti Anselmi de Passione Domini* ('Questions by Saint Anselm about the Lord's Passion'), written in different dialects from Early New High German, Middle Low German, and Middle Dutch. The versions were transcribed in a diplomatic way, and are currently being normalized and annotated with lemma and part of speech. In addition, the versions are being aligned at different levels of granularity (paragraph, sentence, phrase, word). We describe two use cases that profit from the annotations: one use case from historical lexical semantics, the other from historical syntax. We finally sketch further application scenarios from the historico-cultural domain of Digital Humanities.

KEYWORDS: Parallel corpus, Early New High German, lexical semantics, extraposition.

Stefanie Dipper and Simone Schultz-Balluff 2013. The *Anselm Corpus*: Methods and perspectives of a parallel aligned corpus. *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013.* NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87: 27–42.

1 Introduction[1]

This paper deals with the *Anselm Corpus*, a corpus consisting of more than 50 texts of the medieval tract *Interrogatio Sancti Anselmi de Passione Domini* ('Questions by Saint Anselm about the Lord's Passion'). The corpus is being created and annotated in the context of two cooperating projects from linguistics and German medieval studies at Ruhr-University Bochum (Schultz-Balluff and Dipper, 2013).[2]

The different texts are not just one-to-one copies from some source(s) but show considerable variation and, at least in parts, seem to be independent creations.[3] As a consequence, we treat all texts equally, in contrast to most other historical text editions. One of the project goals is a digital edition which gives equal access to all texts of the corpus. Users of the edition will be able to search for important concepts, such as the Last Supper, and compare the different terms used for this concept in the different texts. The edition will also support linguistically-motivated queries, e.g. for investigating the positions of verb arguments or the relative order of auxiliary–verb sequences.

This paper includes a description of the corpus and its annotations. The main focus, however, is on illustrating how this data can be exploited for different kinds of research questions. Based on a small passage, we illustrate research questions from two different areas. One research question concerns historical lexical semantics (the vocabulary), and investigates the different terms used for the concept 'Last Supper', and their temporal and regional distributions, and strategies of conceptualization. The second research question concerns historical syntax, and deals with the distribution of complements and adjuncts. In addition to these research questions from the linguistic domain, we sketch further application scenarios from the historico-cultural domain of Digital Humanities.

The paper is structured as follows. In Sec. 2, we address issues related to editions of historical texts. In Sec. 3, we describe the corpus and its annotations. Sec. 4 and 5 present the two linguistic research questions (use cases) in detail, followed by a sketch of further application scenarios in Sec. 6. Sec. 7 presents an outlook.

2 Text editions

German Medieval Studies usually have a focus on general philological rather than purely-linguistic issues. Since the 19th century, the main goal was on reconstructing the (lost) original source of a historical text. Hence, there is a long tradition in Germany for edition philologists to concentrate on the oldest witnesses of a text. Due to a paradigm shift in the late 20th century, which was initiated by the schools of *New Philology* and *New Historicism*, German Medieval Studies started to look more and more at texts other than the courtly minnesong and epic poetry. Moreover, they were now increasingly interested in the complete tradition of a text, which can span several centuries, and investigated the transmission history and genesis of a text (Bumke, 1996; Quast, 2001). The former premise—"the older the more valuable"—was deprioritized (for an overview of the history of German edition philology, see Bein (1995)).

[1]The research reported here was financed by Deutsche Forschungsgemeinschaft (DFG), Grants DI 1558/4-1 and SCHU 2524/1-1. We would like to thank the anonymous reviewers and the participants of two workshops for helpful comments.

[2]Project URLS: `http://www.ruhr-uni-bochum.de/schultz-balluff/sanktanselmus.html` and `http://www.linguistics.ruhr-uni-bochum.de/anselm`.

[3]Nevertheless, we call the corpus a *parallel corpus* because we treat the texts as parallel texts, by aligning correspondent passages across the texts, see Sec. 3.

Despite the paradigm shift, the goal of digitizing (i.e. transcribing) complete traditions has not yet been achieved because it is a highly time-consuming task.

There is a fairly complete online synopsis of the *Nibelungenlied* ('Song of the Nibelungs'), provided by the University of Vienna[4], which comprises most of the complete and some of the fragmentary manuscripts. However, this project does not aim at diplomatic transcriptions but presents normalized versions, to facilitate readability. For instance, the transcriptions standardize capitalization and the character pairs <i>/<j> and <u>/<v>, and omit many diacritics and some superposed characters. As a consequence, the transcriptions are not suitable for linguistic research.

Another long-term enterprise at the University of Bern[5] aims at digitizing the complete tradition of the epic *Parzival* by Wolfram von Eschenbach. Sample online synopses present fragments of the epic. One of the project goals is to apply methods from *New Phylogeny* to determine stemmatic interrelations.

Other similar projects have just started. For instance, a project at the University of Cambridge will publish a complete edition of the *Kaiserchronik* ('Chronicle of the Emperors').

Our project follows the new paradigm in that it aims at editing all available German and Dutch witnesses of the Anselm text. It goes beyond the above-mentioned projects, though, in that we further enrich the texts by linguistic annotations, at the morphological, syntactic and semantic level, and plan to align the witnesses at different levels: at the level of sections, sentences, phrases and characters. This will enable us to perform comparative linguistic investigations, as is illustrated in this paper.

3 The corpus

Interrogatio Sancti Anselmi de Passione Domini is a tract of the passion, in the form of a dialogue. St. Anselm fasts and prays and implores Virgin Mary to reveal the events of the passion. She finally appears to him and grants his wish. He then starts asking questions, which she answers, about the Passion of Christ, beginning with the Last Supper and ending with the entombment.

The handwritten and printed documents date from the 14th–16th century and are written in a great variety of dialects in Early New High German (ENHG), i.e. Upper, Central, and Low German, and in Middle Dutch. There are 50 texts in total, with an average length of 6,000 tokens. The texts have not yet been investigated in research to any mentionable extent.

3.1 Versions

The tract has been preserved in different versions, which put emphasis on different aspects of the narration, e.g. focussing on Jesus' sufferings or on Mary in her role as the Mother of Sorrows.[6] Based on the different foci and other general properties, the 50 texts can be grouped in 3 different versions: (i) verse versions ("V"); (ii) short prose versions ("PS"); (iii) long prose versions ("PL").[7] They differ with regard to content and distribution:

[4] `http://germanistik.univie.ac.at/index.php?id=14531`

[5] `http://www.parzival.unibe.ch`

[6] See Schiewer (2005) for discussions and criteria of the concepts *version, copy, editing*, and *edition*.

[7] Note that the terms "short" and "long" do not refer to the texts' length but to the texts' content. Long versions are those that add certain specific details to the basic content.

- Verse versions (V) focus on Christ's sufferings whereas prose versions (P) focus on the sorrows of Mary.
- Since verses are written in rhyme, V-versions are rather homogeneous. In contrast, prose as a less formal text form promotes extending the basic content in various ways.
- The opening of the V-versions is very detailed and emphasizes Anselm's scariness and emotions at the moment of Mary's appearance. It includes a justification by Anselm for invoking Mary.

  PS-versions only contain the basic content; details such as Mount Sion or the Golden Gate are mentioned only in PL-versions. PL-versions also often address practical issues worth knowing, such as: What exactly are the "Ismaelitic pennies"? Why can't Mary be alone in the streets after nightfall? How big were the nails used to crucify Jesus?
- V-versions have been preserved from the north and center of Germany, PS-versions from the north and east parts of the German-speaking countries (including Austria and Switzerland). PL-versions stem from the central and southern parts. Only eastern regions produced texts of different versions.

Anselm's first question Both use cases presented in the next sections focus on Anselm's first actual question, which has been preserved in 44 German and 3 Dutch versions. In this question, Anselm asks Mary to describe the beginning of Jesus' martyrdom. Mary starts by describing the Last Supper and the betrayal of Judas. Depending on the respective version (V, PS, PL), the answer can provide further details, such as elaborate explanations of the "Ismaelitic pennies" (which Judas receives for his betrayal), or it contains supplementary elements, such as the footwashing by Jesus.

Table 1 shows Anselm's first question and the beginning of Mary's answer in a verse version, and short and long prose versions.[8]

3.2 Annotations

Normalization Currently, the texts are being annotated semi-automatically with normalized word forms, by mapping the dialect-specific historical word forms to corresponding word forms from modern German (Bollmann et al., 2011). The semi-automatic mapping first produces a simplified word form, where historic characters are replaced by their modern equivalents (e.g. *ſ* is replaced by *s*) and certain abbreviations are spelt out (e.g. ᵔ becomes *us*). In addition, since capitalization and punctuation marks are used in an inconsistent way in the Anselm texts,

[8] The versions are:

- Oldenburg ("O1"): verse version written in Low German, 2nd half of the 14th century; Landesbibliothek Oldenburg, Cim I 74.
- München ("M9"): short prose version in East Upper German, 15th century; Staatsbibliothek München, Cgm 4701.
- Wien ("W1"): long prose version in East Upper German, early 15th century; Österreichische Nationalbibliothek Wien, Cod. 2969.

In the context of the Anselm Corpus, we defined new text sigla O1, M9, W1, etc., that we use throughout this paper. A complete list of the Anselm sigles can be found at `http://www.ruhr-uni-bochum.de/schultz-balluff/sanktanselmus.html`.

Oldenburg (O1; V)	**München (M9; PS)**	**Wien (W1; PL)**
2 Maria erſt wil ik di vraghen ik bidde dattu mi willeſt ſaghen 3 Wu quam it erſt to den pranghen dat din ſone wart ge vanghen	1 Do fragt anzhelm[us] vnd ſprach 2 O aller liebſtew fraw 3 wie hůb ſich an dez erſten deins liebn̄ chindes marter	1 Sand Anſhelm was von herczn vrö und ſprach 2 ſag mir liebe fraw 3 wie was der anfankch der marter dynes libn chindes
5 Ancelme hore dat ik di ſaghe Dat ſchude an dem guden donerſdaghe Dat he mit ſinen jungheren ſaat Lepliken dat he mit on aat He gaf on ſin vleiſch vnd ok ſin blo͝t Dat he vedder vor vns go͝t. [...] 6 Maria do ſe de rede dreuen Wur was judas do ge bleuen Judas de leip alto hant Dar he de Vorſten der jodden vant	4 Do ſprach ma[r]ia 5 Do mein chind an dem antloz tag daz leczt ezzen het mit ſeinē iung[er]n vnd von dem tiſch gie 6 Do gie iudas zu den iuden piſchofen	4 vnſer fraw ſprach 5 da mein libs chind het geeſſenn mit ſeinen Jungern vor ſeiner marter daz leſt mal und da ſy von tiſch auf ſtunden 6 da gieng Judas ſcarioth zw den furſten der Juden

Translation of the Wien-PL text: '1 St. Anselm was very glad and said: 2 Tell me, dear woman, 3 how was the beginning of the martyrdom of your dear child. 4 Our woman said: 5 As my dear child had eaten the Last Supper with his disciples before his martyrdom, and as they left the table, 6 Judas Iscariot went to the high priests of the Jews.'

Table 1: Anselm's first question and the beginning of Mary's answer. Excerpts from a verse version (left column), a short prose version (central column), and a long prose version (right column). Corresponding clauses in the parallel texts are annotated by the same numbers; the PL text serves as the reference text.

all letters are lower-cased and punctuation marks are removed. Next, a cascade of mapping steps are applied to create normalized (= modern) word forms (see below). Table 2 shows a fragment from an Anselm text from East Central German and its simplified and normalized equivalents.

ENHG-orig	ENHG-simpl	NORM		POS	Morph	Lemma
Do	*do*	*da*	then	ADV	–	da
ſprach	*sprach*	*sprach*	said	VVFIN	1.Sg.Past.Ind	sprechen
ſente	*sente*	*sankt*	Saint	NE	Nom.Sg.Masc	Sankt
anſhelm◌̄	*anshelmus*	*anselm*	Anselm	NE	Nom.Sg.Masc	Anselm

Table 2: A fragment from an Anselm text from Early New High German (ENHG). Column ENHG-orig shows the original word form, ENHG-simpl its simplified version and NORM the manually-normalized (modern) equivalent. Further annotations include part of speech (POS), morphology, and lemma.

The cascade of mapping steps (see Bollmann (2012) for details) starts with a word-list mapper that makes use of a lexicon of historical–modern word pairs. Second, word forms not covered by the lexicon are input to a set of character rewrite rules. The rewrite rules are derived from a

set of manually-normalized word forms. Ex. (1) displays a rewrite rule that was derived from the first word pair *do–da* from Table 2.

(1) `o → a / d _ #`
('o' is replaced by 'a' between 'd' and the right word boundary ('#'))

Finally, weighted Levenshtein distance is used to map remaining historical word forms to their "closest" modern counterparts. The edit operations map sequences of up to three characters of historical to modern word forms. Weights for the edit operations are derived from the manually-normalized word forms.

Using small training corpora of 500 manually-normalized word forms, the cascade results in accuracies between 60.71% and 69.34%, depending on the individual text (see Bollmann (2012)).

POS and morphological tagging, lemmatization Normalizing the word forms results in a text that is already close to modern German. It is therefore possible to exploit POS and morphological taggers that have been trained on modern German corpora. Since the normalization procedure only produces lower-case letters and no punctuation marks, Bollmann (2012) retrained the RFTagger (Schmid and Laws, 2008) on modified versions of the TIGER corpus (Brants et al., 2004) and the TüBa-D/Z (Hinrichs et al., 2004) where all letters have been lower-cased and punctuation marks have been removed. Trained on the original data, the RFTagger achieves an accuracy of 96.85%, as compared to 95.74% when trained on the modified data. Removing punctuation especially affects the tagging of relative pronouns, which are often confused with the definite article (Bollmann, 2012).

Applying the RFTagger to the Anselm corpus, the tagger currently achieve accuracies of around 87% on manually-normalized texts and around 76% on automatically-normalized texts (Bollmann, 2012). One reason for the decrease in accuracy, as compared to tagging modern data, is the fact that the training data — which contains newspaper texts — differ considerably from the Anselm texts. For instance, the Anselm corpus contains direct speech (Anselm addressing Mary and vice versa), which frequently involves imperatives, as in *nu hor anſhelmus* 'now listen, Anselm'. In addition, many personal pronouns occur never or only rarely in the training data.

We further plan to annotate morphology and lemma information, see Table 2, as well as syntactic information.

Key words Important concepts and key words, such as *Abendmahl* 'Last Supper', *Karfreitag* 'Good Friday', will be marked in all texts, to facilitate research on these concepts.

Alignment Moreover, corresponding passages are being aligned semi-automatically across the parallel texts (Petran, 2012a). The fact that the texts do not contain punctuation marks to indicate (modern) sentence boundaries poses a special challenge (Petran, 2012b). However, automatic alignment can profit from the normalized word forms.

Semi-automatic alignments are being created at a high level between related questions asked by Anselm, and at a more fine-grained level between corresponding clauses (see the numbering in the fragments in Table 1), phrases (see Table 3) or words. The alignments represent the core annotation of our corpus. They support comparative investigations of the various texts and versions that are part of our corpus, as is illustrated in the next sections.

1. Wien (W1)	2. Halle (H1)	3. München (M4)	4. Karlsruhe (Ka1)
da mein libs chind	do mein lieber son iheſuſz	do mei kint	Do min kint
het geeſſenn	daſz nachtmal	mit ſeinē iungern̄	**hatte gezen**
mit ſeinen Jnngern	mit ſienen iüngern	**het ge eſſen**	daz ivngeſte maz
vor ſeiner marter	*am heiligen grün dornſtage*	*vor ſeiner marter*	mit sinen ivng'n
daz leſt mal	**geſſen hatte**	das iüngſt eſſen	*vor sinꞋ martꞋ*

Table 3: Phrase alignments between four PL texts. Aligned constituents which correspond to one another are highlighted in the same fashion.

Note that the annotations that the comparative investigations are based on are not yet completed at the time of writing. Hence, for the purposes of the pilot studies presented here we created the necessary annotations manually.

4 Use case I: the term and concept *Last Supper*

This section investigates the terms used for the 'Last Supper', by looking at temporal, regional, and version-related distributions and variance (cf. Besch (1967)), and by analyzing strategies of conceptualization (cf., e.g., Busse et al. (1994)). In contrast to the studies by Besch (1967), which focus on High German dialects, the Anselm Corpus also contains texts from Low German and Dutch.

The term *Abendmahl* 'Last Supper', denoting the last meal of Jesus and his disciples, is established in church language only in the early 16th century, heavily influenced by Martin Luther. Prior to that, different, but unambiguous terms had been in use to denote Jesus' farewell dinner. In the following, the different verbalizations of this concept in selected versions of the Anselm text are analyzed, focussing on their temporal and spatial distribution in relation to the three versions V, PS, and PL. The distributions of the variants are displayed in Table 4. The table displays the regions according to their actual locations, starting with Alemanic (1a) and Bavarian (1b) at the bottom (= in the south), and ending with North Low German (5) on the top (= the north).

- The table shows that in almost all **verse versions** (V, in the north: 3a–5), the Last Supper is simply described as the fact that Jesus *mit en at* 'ate with them', i.e. with his disciples.
- In the P-versions, rather fixed (but different) phrases are used. German **short prose versions** (PS, east: 1b–4b)[9] mainly use two terms: combinations of *lezt* 'last' plus *essen* 'meal' in regions 1b (Bavarian) and 2b (East Franconian), and *abent* 'evening' plus *essen* 'meal' in region 1b.

[9] PS-version D4 is from region 2a (west) and seems to be singular in several aspects.

5. North Low German

14 –
15 V *meth em ath* (f1), *mit en at* (Kh1), *myt en at* (Arnd1494)
16 V *mit en ath* (Arnd1521)

4a. Middle Dutch

14 PS *auontmael* (Am1)
15 PS *auont mael* (Le1)
16 PS *auont maeltijt* (Berntsz1523)

4b. West/Eastphalian

14 V *mit em at* (D1), *mit em at* (D2), *mit on aat* (O1)
PS *auent ſpiſe* (Wo1)
15 –
16 –

3a. Ripuarian

14 –
15 V *mit yn as* (KoeldÄ1492), *mit yn as* (KoeldJ1499)
16 V *myt yn as* (Neuss1500), *mit yn as* (Neuss1509), *mit yn as* (Neuss1514), *myt yn as* (Neuss1514/17)

3b. East Central German

14 –
15 V *abint eſſin* (D3)
PS *obent brot* (B1)
PL *nachtmal* (H1)
16 –

2a. Rhenish Franconian

14 –
15 PS *nach mal* (D4)
PL *iungſte maſze/abend eſzen* (B2), *Jungeſte was* (St1)
16 –

2b. East Franconian

14 –
15 PS *letzt eſſen* (Ba1), *leczt eſſen* (Ba2), *leczt eſſen* (N2), *leczſt eſſe* (N3)
PL *iüngſt eſſen* (M4), *oſterlamp* (We1)
16 –

1a. Alemanic

14 –
15 PL *Iung maſz* (Be1), *ivngeſte maz* (Ka1), *Iungſt maſſ* (Stu1), *iungſt mal* (N4), *iüngſt maſz* (sa1), *iungſt <...>* (Sa1), *iüngst male* (Schau1496/97)
16 PL *nacht mal* (SG1)

1b. Bavarian

14 PL *iungiſt mal* (M1)
15 PS *leczt eſſen* (B3), *leſt eſſen* (Me1), *letz eſſen* (M5), *het geſſen* (M6), *abent eſſen* (M7), *abent eſſen* (M8), *leczt ezzen* (M9), *leczt ezzen* (M10)
PL *jungſte mal* (M2), *iwngiſt was* (M3), *abent eſſn̄* (Sb1), *leſt mal* (W1)
16 PL *iungſten mal* (Hk1)

Table 4: Terms and phrases denoting the “Last Supper”, used in different regions, time spans, and versions.

- Similarly, the short versions from Dutch (4a) all use combinations of *abent* 'evening' plus *mal(tijt)* 'meal(time)'. Interestingly, the compound *abentmal* is already spelled in one word in the 14th century version.
- In the **long prose versions** (PL, south and center: 1–2), the combination of *iungst* 'youngest' plus *mal/maz* 'meal' is predominant, occurring in regions 1a, 1b, and 2a.
- Most other occurrences are singular, e.g. *auent spise* 'evening dish' (4b, PS) or *osterlamp* 'paschal lamb' (2b, PL).

The data shows that term selection depends on the region *and* the type of version in combination. Beyond the dominance of certain terms that we mentioned above, no continuity of terms spanning larger regions or time periods can be observed.

The variance that we observe across versions but also within the prose versions seem to suggest that at that time, no general term had yet been established. Terms used already in the 14th century continue to be used in the 15th and 16th centuries; besides them, new forms and combinations were coined.

In one PL version (B2 from region 2a), the term used for the Last Supper is explicitly addressed, see Ex. (2).

(2) Da myn kint hatte geſzen mit ſynen iungern daz iungſte maſze daz da heiſzet daz abend eſzen
'As my child had eaten with his disciples the youngest meal which is called the evening meal'

Iungſte maſze 'youngest meal' is probably a general term, whereas *abend eſzen* 'evening meal' seems to be a more special term, highlighted by the author. However, *abend eſzen* is not a fixed term as can be seen from the variance observed in regions 1b and 3b in the 15th century.

The unsteadiness of the terms is also reflected by the fact that most instantiations are spelled in two words, and only few "real" compounds can be observed: *auontmal* (Am1, 4a) and *nachtmal* (H1, 3b), from the 14th and late 15th centuries, respectively. Moreover, these compounds reoccur later, but spelled in two words (in Le1, 4a, and SG1, 1a).

It is remarkable that the term that has finally been established in standard German is the term from Middle Dutch.

Strategies of specification As we have seen, the prose versions, short and long, seem to struggle for verbalizing the concept *Last Supper* but do not arrive at a common, "standardized" term. The verse versions follow another strategy: they use the unspecific phrase *dat he mit on aat* 'that he ate with them' but add specific temporal information when this happened: *an dem guden donerſdaghe* 'on the good Thursday' (O1).[10]

A similar specification strategy is also followed by some of the prose versions. 11 PS versions[11] add *an dem antlaz tag* 'on the indulgence day'. *Antlaz tag* in general means 'day of

[10] D3 (written in East Central German) represents a special case: It uses rhymes but otherwise shows characteristics of the prose versions. Especially its vocabulary deviates from the other verse versions. This suggests that D3 should be considered separately from the verse versions, and in connection with the prose versions.

[11] Texts B3, Ba1, Ba2, M5, M7, M8, M9, M10, Me1, N2, N3.

release/indulgence', and it can be used to refer to Holy Thursday in particular. Two PL versions (H1, SG1), which use the term *nachtmal* 'night meal', add the specifications *am heiligen gründornſtage* (H1), *am hailgen grünen donstag* (SG1) 'on the Holy Thursday'.

This data shows that the fact that there is not yet a mandatory agreed-upon term is compensated by specification strategies. We propose that the specific strategy used in a subset of the versions can be used as a defining criterion for the version *verse*.

5 Use case II: constituents in situ and extraposed

In the second use case, we select the first sentence of Mary's first answer and compare its different syntactic realizations in all PL versions of our corpus (20 versions in total). In particular, we investigate the positions of verb arguments and adjuncts. Note that this pilot study has been done by aligning the data manually and assigning syntactic functions and positions by hand. In the long run, we plan to do quantitative investigations that cover the entire corpus, by relying on the semi-automatically assigned POS tags and syntactic annotations, and exploiting the alignments at the phrasal level.

In (modern) German, the *right verbal bracket* indicates the boundary between nominal and prepositional arguments and adjuncts that occur *in situ* (preceding the right verbal bracket) or *extraposed* (following the verbal bracket).

In subordinate clauses, the right verbal bracket is filled by verbal components (verbs and auxiliaries), see Ex. (3a). In main clauses, the finite verb or auxiliary takes the second position after some other constituent, filling the *left verbal bracket* (this construction is called *verb-second*). Further verbal components, such as infinite verb forms, verb particles, can occupy the right verbal bracket, see Ex. (3b).[12] The left and right verbal brackets are underlined in the examples. Constituents occurring in situ are marked by "INS", extraposed constituents by "EX". The examples illustrate that in modern German, arguments such as the subject and object occur in situ, whereas adjuncts can be extraposed (optionally).

(3) a. *als* [$_{INS}$ *Jesus]* [$_{INS}$ *das Abendmahl]* *gegessen hatte* [$_{EX}$ *mit* *seinen Jüngern]*
as Jesus the Last_Supper eaten had with his disciples
'as Jesus had eaten the Last Supper with his disciples'

b. *Jesus* *hatte* [$_{INS}$ *das Abendmahl]* *gegessen* [$_{EX}$ *mit seinen Jüngern]*
Jesus had the Last_Supper eaten with his disciples
'Jesus had eaten the Last Supper with his disciples'

The verb-second pattern can already been observed in Old High German, in addition to verb-first, verb-third, and verb-final patterns. In Middle High German, the verb-second pattern has been established as the common structure of main clauses. Verb-final patterns in subordinate clauses are predominant from the earliest stages on. However, as can be seen from Ex. (3a), extraposed constituents can occur after the final verb.

It is well known that arguments and adjuncts occurred in extraposed positions much more frequently in older language stages than nowadays.

Based on data from Gothic, Old English, and different stages from German, Behaghel (1932) shows that short constituents, consisting of one word, predominantly occur in situ, whereas

[12] For a description of the German sentence structure, e.g. see Höhle (1986).

long, "heavy" constituents, e.g. constituents involving coordination, tend to be extraposed. Ebert (1986) examines two texts from the 14th century and finds that around 20% of subordinate clauses contain extraposed constituents, predominantly PPs, but also NP complements.[13]

In the 17th century, the sentence-final position of the verb in subordinate clauses has been established in standard language (Behaghel, 1932, p. 133). That is, since that time, extraposition is limited to clausal arguments and PP adjuncts.

Ex. (4), taken from Behaghel (1932, p. 132), shows an example from Martin Luther with an extraposed object. This construction would be highly marked in modern standard German.

(4) *wenn du erkenntest* $[_{EX}$ *die Gabe Gottes und wer der ist, der zu dir sagt, gib mir trinken]*
'if you knew the gift of God and who it is that asks you for a drink' (John 4:10)

The presentation above shows that in Early New High German, extraposition is still applied to a range of arguments. Hence, it is interesting to investigate the amount of extraposition and the type of arguments that are extraposed in the different Anselm texts. To do this, we analyse the first sentence of Mary's first answer in detail, see Table 5. The table displays the W1-text in the first column, organized by constituents, a translation in the second column, each constituent's function in the third column, and its position in the forth column. The sentence consists of two subordinate clauses, followed by the main clause.

As can be seen from the table, the three subjects occur in situ. The remaining constituents of the first subordinate clause are extraposed, in contrast to the constituents of the second subordinate clause. The positions of the main clause constituents cannot be determined in this example because the right verbal bracket is not filled (but see below).

The distribution of the constituents, as realized in this text, is in fact the "default" distribution, which shows up in 11 of the 20 PL texts.[14] In three texts,[15] all constituents occur in situ. Interestingly, these texts share another unique feature: the verbal components of the first clause (line 3) show the modern order *verb participle* > *finite auxiliary*, e.g. *geffen hett* 'eaten had', in contrast to all other texts.

In 14 texts, the NP-object of the first clause (line 6) occurs after the PP-adjuncts (lines 4 and 5). In five texts,[16] the NP-object occurs in front of the PP-adjuncts, and in one text,[17] it occurs between both PPs.

In three texts,[18] the right verbal bracket of the main clause is filled by a verb particle. In these cases, the locative PP, which denotes the goal of the movement (line 15), is extraposed, see Ex. (5).

[13]The terms *in situ* and *extraposed* suggest that one of the positions is the "original", unmarked one, while the other is a secondary position, derived from the first, e.g., by a relation called "extraposition". For modern German, the unmarked positions of NP and PP constituents are clearly in front of the rigth verbal bracket, and positions behind the right verbal bracket are exceptional. In former stages of German, however, the situation is not as clear. Hence, the reader is asked to interpret the terms *in situ* and *extraposed* as referring to pre- and postverbal positions, without implications about the actual analysis.

[14]From region 1a: N4, Stu1, Schau1496/97; region 1b: M1, M2, M3, W1, Hk1; region 2a: B2, St1; region 2b: We1.

[15]From region 1a: SG1; region 1b: Sb1; region 3b: H1.

[16]Region 1a: Be1, Ka1, sa1, SG1; region 3b: H1.

[17]Region 1a: Sa1.

[18]All from region 1a: Be1, Ka1, Sa1

	Wien (“W1”, PL)		Function	Position	Clause
1	*da*	as	subord		
2	*mein libs chind*	my dear child	NP-subj	INS	Subord 1 (1–6)
3	*het geeſſenn*	has eaten	verb	right VB	
4	*mit ſeinen Jungern*	with his disciples	PP-adjunct	EX	
5	*vor ſeiner marter*	before his martyrdom	PP-adjunct	EX	
6	*daz left mal*	the Last Supper	NP-obj	EX	
7	*und*	and	coord		
8	*da*	as	subord		
9	*ſy*	they	NP-subj	INS	Subord 2 (8–11)
10	*von tiſch*	off table	PP-adjunct	INS	
11	*auf ſtunden*	up stood	verb	right VB	
12	*da*	then	adverb		
13	*gieng*	went	verb	left VB	Main (12–15)
14	*Judas ſcarioth*	Judas Iscariot	NP-subj	?	
15	*zw den furſten der Juden*	to the princes of_the Jews	PP-goal	?	

Table 5: The beginning of Mary’s first answer: ‘As my dear child had eaten the Last Supper with his disciples before his martyrdom, and as they left the table, Judas Iscariot went to the high priests of the Jews’. INS: in situ, EX: extraposed, VB: verbal bracket.

(5) *Do* *giench* *[INS* *ivdas* *ſcarioth]* *vz.* *[EX* *zv* *den* *fv̈rſten* *d ϟ* *ivden].*
then went Judas Iscariot out to the princes of_the Jews
‘Then Judas went out to the high priests of the Jews’

Further differences between the texts include: absence of the second subordinating conjunction (line 8); absence of the subject in the second subordinate clause (line 9).

To sum up the findings of this small comparison, we have seen that extraposition of the object NP seems to be the unmarked case, in contrast to modern German. Ignoring the case of subject NPs (which seem to be extraposed only rarely) and unclear positions, the numbers of constituents in situ vs. extraposed are almost equal in the default order: two PPs and one object NP occur extraposed, two PPs are in situ. Our small study seems to indicate that the Anselm texts exhibit considerably more extraposed constituents than the texts examined by Ebert (1986), possibly due to the fact that the Anselm texts contain direct speech. Fully-annotated texts will allow us to investigate such questions to a greater extent and in more detail.

6 Further application scenarios

The use cases presented in Sec. 4 and 5 dealt with linguistic issues. In this section, we want to sketch the use of the Anselm Corpus and its annotations for research in other fields of the humanities, such as reception history or narratology. In particular, we show that Mary and Jesus are addressed and referenced in different ways in the individual texts. Taken together with other characteristics of the texts, we can deduce from the different ways of reference that the texts were composed for different groups of recipients.

Forms of address The forms of address for **Mary** vary considerably between the different texts. In the prose versions, Anselm addresses Mary by forms indicating devotion: *leue vrowe* 'dear woman' (Wo1), or *aller liebſtew fraw* 'most dearest woman' (M9). When talking about Jesus, Anselm emphasizes Mary's role of the mother: *deins liben kindes* 'your dear child' (N3). This culminates in the description of Mary as the mother of all humans: *liebe mutt[er]* 'dear mother' (form used by Anselm, H1) or as the Mother of God: *Mutt[er] gotteſz* (form used by the narrator, H1). Finally, the narrators of the prose texts involve the recipient, e.g. by phrases such as *vnſer fraw* 'our woman' (e.g. W1) or *vnſe liebe frauwe* 'our dear women' (e.g. B2).

In the verse versions, the relation to Mary remains more reserved. She is addressed exclusively by her name. Still, her mother role is present in that Jesus is referenced by *din ſon* 'your son'.

When talking about **Jesus**, Mary refers to him by *min kint* 'my child' (Be1), *mein libs kint* 'my dear child' (W1) or even *min alre lieffte kint* 'my most dearest child' (Am1). In the verse versions, Jesus is predominantly referred to by the personal pronoun.

In short, the prose versions emphasize the relation mother–son, described from the point of view of the sorrowing mother, and establish a mother relation between Mary and the recipient of the text. The personal relationship is intensified by elaborate passages of lamentation. The idea of compassion and the role of Mary as Mater Dolorosa plays a stronger role in the prose than in the verse versions, which remain more distant in general.

Incipit and explicit The forms of the incipits and explicits (the opening and closing words of the texts) also show that the texts were composed and intended for different groups of users. For instance, the explicit of the text Wo1 begins with the words *Swe dit het lest de vordenet seszdusentseshundertvndsesvndsestindich iar afflates* 'whoever has read this deserves 6,666 years of indulgence'. This is followed by a compact depiction of Jesus' sufferings. The text ends with the words *des danke ic di here* 'I thank you for that, lord', emphasizing the prayerlike style of the explicit. One can suppose, therefore, that the text served purposes of intensive internalization and spiritualization, and could have been used in a monastic context. In fact, the text is part of a larger manuscript consisting of multiple individual texts; the composition of this manuscript also supports the presumption that it was used in a spiritual, religious context, presumably in a nunnery in the Eastphalian region (Schultz-Balluff, 2013).

As we have just seen, the recipient of the text Wo1 is actively involved. The explicit of the text Kh1 shows a different picture. The text ends with a plea to God to send peace to the human beings, and states that everyone who would not enjoy the tract should remain a fool forever: *Dyt is sunte ansylmus vrage/ weme se nicht en behage/ De blyue en schalk al syne daghe* 'This is St. Anselm's question. Who does not like it remains a fool for the rest of his life'.

According to some of the texts, reading *St Anselms Question's about the Lord's Passion* can also be helpful for the everyday life. For instance, it facilitates childbirth (M8): *den frawn̄ die do ſchwanger ſein vnd ſwarlig kinder gepern den iſt dyſz büchlen alſz nůz als ob ſy andere ding theten dy den frawn̄ hylff geben* 'For pregnant women who have difficulties in bearing children this book is beneficial just like other things that might help women in such a situation'. Similarly, reading the tract can protect the house against harm caused by water or other disaster (M8): *in welgē hauſz das buch mit andacht wurt geleſen vnd in welgē haüſz es iſt dem ſelbn̄ hauſz kan kain waſſer ader kain vngehewr geſchadn̄* 'Houses in which this book is read with devotion, and houses in which it is present, cannot be harmed by water or disaster'.

These examples illustrate that the tract in its different forms served different kinds of recipients:

in monasteries and convents, other clerical circles, and also in the secular part of society.

To sum up, the forms of address, style, and differences in content allow us to draw conclusions with regard to the context of use of individual texts or versions, and to the image of Mary and the envisaged recipients.

7 Conclusion

In this paper, we presented a parallel corpus of texts from Early New High German and Middle Low German. We argued that alignments at different levels (question–answer pairs, sentences, phrases, words) can support comparative investigations in different areas. This was illustrated by different use cases from historical lexical semantics (comparing terms used for the Last Supper), historical syntax (comparing the distribution of constituents), and from a historico-cultural perspective (comparing the ways Mary and Jesus are addressed and referenced, and the specific forms of the incipits and explicits). The last scenarios showed that linguistic annotations can also benefit non-linguistic research in Digital Humanities.

The (manually-created) annotations that were made use of in the pilot studies are:

- Use case 1 ("Last Supper"): alignment at paragraph level (Anselm's questions) and key words
- Use case 2 (extraposition): alignment at chunk/phrase level
- Further application scenario 1 (forms of address): key words
- Further application scenario 2 (incipit and explicit): alignment at paragraph level

We plan to create a digital edition of the entire corpus (Stolz et al., 2007). Users will be able to select texts from the collection and search for specific word forms, parts of speech etc. The query results will be presented in the form of a synopsis, which places aligned passages next to each other.

We think that the alignments can also support semi-automatic creation of a critical apparatus used in a print edition. The variance observed between the three versions (verse, short prose, long prose) suggests that all three versions would be edited. The variance could also lead to considerations whether we actually deal with one or three texts.

The Anselm corpus is especially well suited for a pilot project exploring comparative linguistic research based on an aligned parallel corpus. The Anselm texts are rather short but written in a great variety of dialects. In future work, we would like to apply our method to other texts with many surviving witnesses, such das *Die 24 Alten* from Otto von Passau ('The 24 Elders'; 142 witnesses, among them 11 fragments) or *Unser vrouwen klage* ('Mary's lamentations'; 26 witnesses, 4 fragments). Next, we would like to look at other types of texts, e.g. pharmacopoeias, which have a long and broad tradition throughout the entire Middle Ages, e.g. *Bartholomäus* (with 32 witnesses, 6 fragments).

The long-term goal would be to analyze long texts using our methods, such as *Parzival, the Song of the Nibelungs* or *the Chronicle of the Emperors*.

References

Behaghel, O. (1932). *Deutsche Syntax. Eine geschichtliche Darstellung. Band IV: Wortstellung. Periodenbau*. Winter, Heidelberg.

Bein, T., editor (1995). *Altgermanistische Editionswissenschaft*. Peter Lang, Frankfurt/Main, New York.

Besch, W. (1967). *Sprachlandschaften und Sprachausgleich im 15. Jahrhundert. Studien zur Erforschung der spätmittelhochdeutschen Schreibdialekte und zur Entstehung der neuhochdeutschen Schriftsprache*. Francke, München.

Bollmann, M. (2012). Automatic normalization for linguistic annotation of historical language data. Master's thesis, Ruhr-Universität Bochum.

Bollmann, M., Petran, F., and Dipper, S. (2011). Applying rule-based normalization to different types of historical texts — an evaluation. In Vetulani, Z., editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 339–344, Poznan, Poland.

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Bumke, J. (1996). Der unfeste Text. Überlegungen zur Überlieferungegeschichte und Textkritik der höfischen Epik im 13. Jahrhundert. In Müller, J.-D., editor, *Aufführung und Schrift in Mittelalter und Früher Neuzeit*. Metzler, Stuttgart, Weimar.

Busse, D., Hermanns, F., and Teubert, W., editors (1994). *Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik*. Westdeutscher Verlag, Opladen.

Ebert, R. P. (1986). *Historische Syntax des Deutschen II: 1300-1750*. Peter Lang, Frankfurt.

Hinrichs, E., Kübler, S., Naumann, K., Telljohann, H., and Trushkina, J. (2004). Recent developments in linguistic annotations of the TüBa-D/Z Treebank. In *Proceedings of TLT 2004*, Tübingen, Germany.

Höhle, T. (1986). Der Begriff 'Mittelfeld'. Anmerkungen zur Theorie der topologischen Felder. In Schöne, A., editor, *Kontroversen, neue und alte. Akten des 7. Internationalen Germanistenkongresses Göttingen 1985*, pages 329–340. Niemeyer, Tübingen.

Petran, F. (2012a). Aligning the un-alignable — a pilot study using a noisy corpus of nonstandardized, semi-parallel texts. In Gelbkuh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 2. Springer, Berlin, Heidelberg.

Petran, F. (2012b). Studies for segmentation of historical texts: Sentences or chunks? In *Proceedings of the TLT-Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), 2012*, Lisbon, Portugal.

Quast, B. (2001). Der feste Text. Beobachtungen zur Beweglichkeit des Textes aus Sicht der Produzenten. In Peters, U., editor, *Text und Kultur. Mittelalterliche Literatur 1150–1450*. Metzler, Stuttgart, Weimar.

Schiewer, H. J. (2005). Fassung, Bearbeitung, Version und Edition. In Schubert, M. J., editor, *Deutsche Texte des Mittelalters zwischen Handschriftennähe und Rekonstruktion. Berliner Fachtagung 1.-3. April 2004*. de Gruyter, Tübingen.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, Manchester, UK.

Schultz-Balluff, S. (2013). Auf dem Wandbord einer Nonne — Ein Passionstraktat in täglichem Gebrauch. In *Rosenkränze und Seelengärten. Bildung und Frömmigkeit in niedersächsischen Frauenklöstern*, Ausstellungskataloge der Herzog August Bibliothek 95, pages 147–155. Herzog August Bibliothek Wolfenbüttel, Wiesbaden.

Schultz-Balluff, S. and Dipper, S. (2013). 'St. Anselmi Fragen an Maria' — Schritte zu einer (digitalen) Erschließung, Auswertung und Edition der gesamten deutschsprachigen Überlieferung (14.–16. Jh.). In Bohnenkamp-Renken, A., editor, *Medienwandel/Medienwechsel in der Editionswissenschaft*, Beihefte zu editio, pages 173–191. Berlin, Boston: de Gruyter.

Stolz, M., Lucas, M., and Loop, J., editors (2007). *Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien. Eine Standortbestimmung*. germanistik.ch.

Finite-state Relations Between Two Historically Closely Related Languages

Kimmo Koskenniemi

University of Helsinki, Finland

kimmo.koskenniemi@helsinki.fi,

ABSTRACT

Regular correspondences between historically related languages can be modelled using finite-state transducers (FST). A new method is presented by demonstrating it with a bidirectional experiment between Finnish and Estonian. An artificial representation (resembling a proto-language) is established between two related languages. This representation, AFE (Aligned Finnish-Estonian) is based on the letter by letter alignment of the two languages and uses mechanically constructed morphophonemes which represent the corresponding characters. By describing the constraints of this AFE using two-level rules, one may construct useful mappings between the languages. In this way, the badly ambiguous FSTs from Finnish and Estonian to AFE can be composed into a practically unambiguous transducer from Finnish to Estonian. The inverse mapping from Estonian to Finnish is mildly ambiguous. Steps according to the proposed method could be repeated as such with dialectal or older written texts. Choosing a set of model words, aligning them, recording the mechanical correspondences and designing rules for the constraints could be done with a limited effort. For the purposes of indexing and searching, the mild ambiguity may be tolerable as such. The ambiguity can be further reduced by composing the resulting FST with a speller or morphological analyser of the standard language.

KEYWORDS: Finite-State Transducers, Historical Linguistics, HFST, Two-Level Morphology, Foma.

Kimmo Koskenniemi 2013. Finite-state relations between two historically closely related languages. *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013.* NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87: 43–53.

1 Introduction

In historical linguistics one studies two or more languages which are assumed to be related. Among other things, the scholar collects cognate words, postulates the sounds in the proto-language and establishes regular correspondences or so called sound laws according to which the known languages can be derived from the proto-language. For a more detailed description of the comparative method, see (Campbell, 2004, pp. 127–147). When establishing the proto-language and the correspondences, the scholar both compares the known languages, i.e. uses the external evidence and studies the sound alternations within single languages, i.e. uses the internal evidence. Different scholars may disagree upon the details of proto-languages and often there is even a dispute whether a common ancestor can be established at all, cf. (Campbell, 2004, pp. 345–346).

Finite-state rewriting systems like the XFST (Beesley and Karttunen, 2003) and its open source counterpart FOMA (Hulden, 2009) provide basic tools for expressing and implementing the regular sound relations. In addition, there is an open source toolkit HFST (Lindén et al., 2011) based on FOMA, SFST (Schmid, 2005) and OpenFST (Allauzen et al., 2007). HFST provides a programming interface and command line tools including HFST-TWOLC, a compiler for two-level rules. Regular relations are implemented as finite-state transducers (FST) which transform strings to other strings.

In particular, by using FSTs, one may explicitly check whether proposed sound laws produce the correct results or not. Rewrite rules implement more or less directly the phonological rules and a sequence or cascade of such rules can alter the word in the proto-language into a word in the known language.

FSTs have the advantage that the mappings can be readily inverted or combined with other mappings. Thus, once you have designed rules for deriving language A from a proto-language C, you may invert the rules, and see what possible proto-forms any word of A could have. Such an inverted mapping can be composed with the mapping from C to language B to see which words in B could be of the same origin as a word in A. The inverted rules, e.g. the mapping from C to A, tend to be severely ambiguous. Some description of the morphophonological constraints, e.g. of the proto-language C are necessary, in order to reach reasonably bidirectional mappings.

The source of ambiguity is about the same in the traditional historical sound relations and in the approach presented here. This paper carries out a small scale experiment using Estonian and Finnish cognate words. Instead of using the proto-language, an artificial language through alignment is used. The expressing of the redundancy of the proto or artificial language is the way to produce usable reverse relations.

The present work is not statistically oriented. The reader is advised to consult e.g. (Bouchard-Côté et al., 2009; Bouchard-Côté et al., 2013) for approaches which reconstruct proto-languages in an unsupervised manner and construct phylogenetic trees. This work also differs from the data driven approached based on Minimum Description Length (MDL). For more information on such, the reader should consult e.g. the work of Roman Yangarber and his colleagues (Wettig et al., 2012).

The present work relies on the general linguistic knowledge and linguist's intuition. The conceptual framework in this approach differs in certain respects from the widely accepted comparative method. The most notable difference is the use of direct relations instead of the stepwise application of rewrite rules. The use of two-level parallel rules in controlling the

ambiguity makes the individual decisions of the linguist less dependent of each other. Testing of two-level rules is easy because the tools pinpoint the contradicting rule and the point where something in the example does not match the rule.

Even if a linguist does not agree with the framework presented below, she or he may use the method for building practical and useful linguistic modules. In particular, the method may turn out to be useful in describing the relation between dialects or closely related languages. One could also think of extending the method to cope with more realistic proto-language deduction and the study of historical linguistics.

2 Toy Balto-Finnic

This paper reports an experiment with two related languages, Finnish (FI) and Estonian (ET) which are understood to derive from Proto Balto-Finnic (PBF). The paper tries to show how the present tools can be used constructively when one needs to relate two languages or language forms which are close to each other.

Exercise data from (Campbell, 2004, pp.59–61) is used for this experiment. The book gives some 84 carefully selected cognate words in Finnish and Estonian. Some examples of the list which are genetically related words are given in Table 1.

PBF	*FI*	*ET*	*Gloss*
mees	mies	mees	'man'
ikä	ikä	iga	'age'
lehti	lehti	leht	'leaf, sheet'
verkko	verkko	võrk	'net'

Table 1: Sample words: Proto Balto-Finnic, Finnish and Estonian

The written forms of Finnish and Estonian words are used here. Thus some identical sounds are represented using different letters, e.g. **y** and **ü** as well as the Finnish **k** and the Estonian **g**. If one would do serious historical linguistics, one would use e.g. symbols from the International Phonetic Alphabet (IPA)[1] for a phonologically motivated representation of sounds. In the present case, the use of the written forms causes little harm as both languages are written almost phonemically. The method is insensitive to the slight differences in the orthographic conventions.

3 Aligning the words

The Proto Balto-Finnic form is also given in the book, but it is not used in the experiment. Instead, the *Finnish and Estonian forms are aligned letter by letter with each other* using general linguistic knowledge. Vowels may match vowels and consonants consonants. Semivowels match semivowels but may even match vowels or consonants. Estonian words tend to be shorter than the Finnish ones, so some letters in one language may correspond to zero in the other. The above four words would give the following alignment:

```
m ie e s          (*mees)
i kg äa           (*ikä)
l e h t iØ        (*lehti)
v eõ r k kØ oØ    (*verkko)
```

[1] See e.g. `http://en.wikipedia.org/wiki/International_Phonetic_Alphabet`

Single letters in this alignment indicate that identical letters correspond to each other in that position. A pair, e.g. **ää** indicates that **ä** in Finnish corresponds to **a** in Estonian. This alignment can be done in a fairly objective manner so that almost anybody (or at least any linguist) would arrive at the same result. Matching vowels with vowels and consonants with consonants is nearly sufficient as a criterion. In addition, semivowels e.g. **j** may match either consonants or vowels. Deletion and epenthesis must, of course be allowed, if needed, e.g. **iØ** when there is an **i** in Finnish but nothing in the Estonian form.[2]

Next, we treat this representation of the aligned words as a substitute of the proto-language and call it Aligned Finnish-Estonian (AFE). The AFE has clearly many more (morpho)phonemes than the conventional PBF. In our case, the AFE contains all information that the PBF does and some additional information. We may map all our example words of the AFE into Finnish, Estonian or even to PBF e.g. by using a parallel replace rule of XFST or FOMA. The mappings are many-to-one because there are more symbols in the AFE than in any of the real languages or in PBF. For example, the mapping from the AFE into Finnish using XFST or FOMA would be:

```
regex [aØ->a, eõ->e, ie->i, iõ->i, iØ->i, ji->j, kg->k, kØ->k,
 ou->o, oõ->o, oØ->o, pb->p, pØ->p, td->t, uo->u, uØ->u,
 yü->y, yö->y, yØ->y, äa->ä, äØ->ä, nØ->n, Øa->0, Øü->0, Øõ->0] ;
```

Here all single letters are unchanged and the second letter of two-letter symbols is dropped. The transformation from the AFE into Estonian drops, respectively, the first letter.

The mapping from the AFE to PBF takes mostly the first of the two letters with some exceptions:[3]

```
regex [aØ->a, eõ->e, ie->e, iõ->e, iØ->i, ji->j, kg->k, kØ->k,
 ou->o, oõ->o, oØ->o, pb->p, pØ->p, td->t, uo->o, uØ->u,
 yü->y, yö->ö, yØ->y, äa->ä, äØ->ä, nØ->n, Øa->0, Øü->0, Øõ->0] ;
```

This mapping appears to be surprisingly simple considering the fact that we did not use our knowledge about PBF when constructing the AFE representation. The mapping is simple perhaps because the two languages are so close to each other and because the examples we included did not cover the more complex cases. If we can do the same with a comprehensive set of examples, we might conclude that such a *proto-language only has features which are present* in the known languages. Nothing has then been entered into the proto-language that was not present (in some form) in either of the languages.

4 Internal regularities of the AFE

Writing the rules which transform this kind of a pseudo proto-language AFE into Finnish, Estonian and PBF is trivial, but not very useful as such. The inverted transformations of these,

[2]At this stage, the alignment was made manually. The author also tried a C program made by Måns Huldén which made the alignment automatically using the method of Gibbs sampling. The automatic alignment was almost identical to the manual on, differing only in three words such as Finnish **kansi** vs. Estonian **kaas** where algorithm aligned **k a n:a s i:Ø** whereas my own alignment was **k a Ø:a n:Ø s i:Ø**. The PhD dissertation of Kondrak gives a good survey of the various methods which have been used in aligning and identifying cognate words in related languages, see (Kondrak, 2002).

[3]The exceptions deal with diphthongs in Finnish which correspond to long vowels in PBF. These include the AFE **iõ ie** (these occur only where there is a diphthong **i e** in Finnish and **e e** in PBF), the AFE **uo** (corresponding to a part of **u o** in Finnish and **o o** in PBF) and the AFE **yö** (a part of **y ö** in Finnish and **ö ö** in PBF).

e.g. from Finnish to AFE, are heavily or infinitely ambiguous unless some constraints are applied. Many letters in Finnish or Estonian have multiple possible counterparts in the AFE. Even worse, a deletion (e.g. in AFE-TO-FI) causes an infinite cycle in the inverted mapping (FI-TO-AFE). Thus the raw composition of FI-TO-AFE and AFE-TO-ET would map the Finnish word to its Estonian cognate and a host of other, unwanted forms.

4.1 Filtering excessive possibilities

In order to make the inverse mappings useful, we describe the *regularities which constrain the AFE*, i.e. in what contexts its (artificial) morphophonemes may occur and where they may not. Combining such a filter with the trivial mappings helps in removing many or most of the unwanted strings. Finnish words could be transformed into Estonian words (assuming that they are related) by a composition of three transducers (where **.o.** stands for composition):

```
FI-TO-AFE .o. AFE-FILTER .o. AFE-TO-ET
```

The first is the inverse transducer of the above FOMA replace rule which takes a Finnish word as input and produces a (possibly infinite) set of AFE strings which AFE-TO-FI would map to the input word. AFE-TO-ET is FST for the parallel FOMA replace rule for Estonian (which maps one AFE string into one Estonian word).

The middle transducer, AFE-FILTER is the key component of the present solution. It uses the alphabet of AFE for its input and output. AFE-FILTER never alters anything. Instead, it removes or forbids sequences which do not conform with the regularities. AFE-FILTER is here implemented using two-level rules see (Koskenniemi, 1983; Karttunen, 1993). To compile the rules into FSTs, HFST-TWOLC, the open source two-level rule compiler was used (Silfverberg and Lindén, 2009). The filter uses two kinds of rules, (1) right-arrow rules (=>) which list the contexts a morphophoneme may only occur, and (2) exclusion rules (/<=) which list contexts where a morphophoneme may not occur.

The rules were written by according to normal linguistic intuition using the AFE form of the example words and studying the distribution of each two-letter morphophoneme in turn. Linguistic judgements were based on a text file containing all example words in their AFE representation. This file was searched using simple regular expressions. On an ordinary Unix or Linux machine, one would use the GREP or EGREP program by searching for interesting morphophonemes and selecting only those lines which contain them. The GNU Emacs has an equivalent command (M-x occur) which was actually used for this purpose.

For determining the constraints concerning Finnish **u**, one looks at the distributions of AFE morphophonemes **u**, **uo** and **uØ**. It is easy to see that **uo** only occurs in the first syllable as a part of a diphthong in Finnish (or a long vowel in Estonian). The **uØ** only occurs at the end of the word where it must be preceded by a double consonant or a long vowel or a diphthong. In this way, constraints for each AFE morphophoneme can be formulated. The constraints are quite independent of each other and there is no ordering among them. All constraints must be respected separately and applying one does not make any of the other constraints less or more applicable. (As opposed to rewrite grammars, two-level grammars have neither bleeding nor feeding.)

No claims are made for the completeness or generality of the rule set produced in this experiment, but it works quite nicely for the examples in Campbell's book. Almost all Finnish example

words map into just one (correct) Estonian word. In the inverse direction, most Estonian example words produce a set of Finnish words so that one member of the output set is the correct one, e.g. Estonian **haav** is mapped into Finnish **haava**, **haavi**, **haavo** and **haavu** where the first in this list happens to be the correct one.

4.2 Rules for the constraints

The experiment implemented the filter using the HFST-TWOLC two-level compiler. Normally, two-level grammars relate the lexical and surface representations in order to describe the morphophonological alternations in the inflection. Here, we used the rule compiler in an unusual way where only one relevant representation, the AFE, was involved. The alphabet consisted of all letters and so called morphophonemes in the AFE.

```
Alphabet
  a b d e f h i j k l m n o p r s t u v y ü ä ö aØ eõ ie iõ iØ ji kg kØ
  ou oõ oØ pb pØ td uo uØ yü yö yØ äa äØ nØ Øa Øü Øõ ;
```

With this alphabet, definitions for vowels **V** and consonants **C** are needed, and a particular phonological environment **Dbl** which appears to control many features in Estonian. If a double consonant precedes immediately or a double vowel precedes the single consonant, then certain stem final consonants are deleted.

```
Sets
  V = a e i o u y ä ö aØ eõ ie iõ iØ ou oõ oØ uo uØ yö yü yØ äa äØ ;
  C = b d f g h j k l m n p r s t v ji kg kØ nØ pb pØ td ;
Definitions
  Dbl = [V V | C | V Z nØ] C (V C) ;
```

Using these sets and definitions, certain sequences are excluded from the AFE representations. E.g. for **u** we have two constraints each of which is expressed with two rules. The first rule restricts the possible contexts where an AFE morphophoneme may occur and the second excludes other alternatives in such a location. The first constraint is for long vowels in Estonian (and in PBF) which in Finnish are diphthongs. The second constraint controls the deletion of the stem final vowel in Estonian.

```
Rules
...
"uo"  uo   =>  .#. C* _ o ;
"~u"  u   /<=  .#. C* _ o ;
"uØ"  uØ   =>  Dbl _ .#. ;
"~u"  u   /<=  Dbl _ .#. ;
...
```

The first rule simply says that the AFE morphophoneme **uo** can occur only if there are only consonants to the left of it and an **o** immediately to the right. The second rule forbids the occurrence of **u** in this context. The third rule tells that the AFE morphophoneme **uØ** can only occur at the end of the word and if it is immediately preceded by something that matches the predetermined expression **Dbl** given above. **Dbl** essentially represents either a double consonant or a double vowel as expressed in the regular expression formalism used in the Xerox XFST.

The authoring of these filtering rules proved to be easier than expected because good tools were available. The HFST-PAIR-TEST program tests the compiled two-level rules. It uses a test file which was derived from the AFE representation of the set of examples using, among other things, the AFE-TO-FI transducer. The pair test program immediately points out discrepancies between the examples and the newly written rules. A few hours of intensive writing, testing and tuning was sufficient.

Then, FI-TO-AFE, AFE-FILTER and AFE-TO-ET were composed into a single FST. The testing of this with the input of Finnish test words gave almost clean results. Some tuning was necessary, especially because there were some deletions in the mapping from AFE to Finnish. Soon the filter enabled almost clean results in the mapping from Finnish to Estonian (88 results out of 84 input words).

The inverse mapping from Estonian to Finnish was ambiguous and produced some 4.5 results per input word. This ambiguity was almost exclusively due to the deletion of the stem final vowels in Estonian.

The experiment presented here is a toy. It could easily be expanded to cover a larger part of the regular sound relations between Finnish and Estonian. The result would be an interesting tool for the linguist or a linguistically oriented language learner rather than a practical translator between these languages. The non-deterministic mapping from Estonian to Finnish could still be made more unambiguous by using a finite-state speller for Finnish, e.g. OMORFI (Pirinen, 2011). Composing the Estonian to Finnish mapping with such a speller FST would exclude most of the wrong results because the random addition of the stem final vowel mostly produces non-words in Finnish.

4.3 Alternative implementation

It is clear that one could have expressed the whole mapping from Finnish to Estonian using just the surface representations of these languages and one single two-level grammar. In this grammar, the alphabet consists of true pairs instead of the atomic two-letter morphophonemes. In this implementation, the declaration of the alphabet lists units like **ä ö a:0 i:õ**. The rules combine the pair of two-level rules needed for one constraint in the previous approach, each into a single rule e.g.:

```
"u:o"  u:o  <=>  .#. C* _ o ;
"u:∅"  u:0  <=>  Dbl _ .#. ;
```

Authoring such rules was not essentially more difficult or time consuming than writing the grammar which checked just the AFE consistency. The present experience suggests that the first approach is somewhat easier to manage. The first approach is also more flexible in cases where one would like to permit certain overlapping contexts for distinct pairs (or morphophonemes). The AFE based filter appears easier to extend or modify allows the linguist to choose different levels of strictness. A particular asset of the first approach is the modularity and symmetry of it. It can be used in additional combinations, e.g. as a component for building Estonian to PBF.

5 Potential applications

The result of the experiment is probably not very useful in other ways than that it extends our understanding and that similar tools may be constructed for more concrete and practical applications. Some applications have been considered, including one related with the Corpus of

Old Finnish texts.[4] The frequency list of the corpus is available on line.[5] Below is a sample of the word frequency list of Old Finnish (OF):

```
75      4116    ioca    0,1201 %
76      4074    händä   0,1189 %
77      3952    hänelle 0,1153 %
...
81      3797    ombi    0,1108 %
82      3787    canssa  0,1105 %
83      3775    sijnä   0,1102 %
```

One could process selected samples of OF and align them with the corresponding words of Modern Standard Finnish (MSF). Proceeding according to the principles and steps above, one would get aligned pairs, corresponding rules, and a transducer which converts MSF to OF and vice versa. This OF-TO-MSF could be combined with a list of possible word forms of MSF, which we can readily get from a Finnish morphological analyser in FST form (Pirinen, 2011). In addition to these components, one would probably want to include a list of exceptions where the OF word is not in a regular relation with the MSF word.

6 Deducing the proto-form

Let us return to the example rules which were discussed in 4.2. The following rules correspond to Finnish **ä**. Let us look, how we can reduce the set of corresponding AFE morphophonemes **äa**, **ä** and **äØ** into a smaller and more realistic set.

```
"äa"  äa   => V C+ _ ;
"~ä"  ä   /<= V C+ _ ;
"äØ"  äØ   => Dbl _ .#. ;
"~äa" äa  /<= Dbl _ .#. ;
```

We can see that the two correspondences (**äa** and **äØ**) allowed by these rules are restricted to clearly defined contexts and that the third one **ä** is allowed elsewhere, i.e. it has no context condition which could be expressed in clear linguistic terms. Therefore, **ä** is a good candidate to represent these three AFE morphophonemes. Suppose that we would change the AFE by reducing **äa** and **äØ** into **ä**. Then we have to replace the simple rewrite rule `äa -> a` (from AFE to Estonian) by rules with a context:

```
ä:a <=> V C+ _ ;
```

Similarly, instead of `äØ -> 0` we would have:

```
ä:Ø <=> Dbl _ .#.
```

After these modifications, the mappings still produces the correct results and we may proceed to make further reductions by joining other AFE morphophonemes. As we join the AFE morphophonemes, we build the mappings by replacing one by one the trivial translation rules of AFE-TO-ET and AFE-TO-FI with rules that correspond to the two-level rules of AFE-FILTER. This process stops when no more simplifications of the AFE morphophonemes can be made. The resulting rule sets correspond to the sound laws according to which the present Finnish and Estonian can be derived out of this newly created candidate for the proto-language.

[4] `http://kaino.kotus.fi/korpus/vks/meta/vks_coll_rdf.xml`
[5] `http://kaino.kotus.fi/sanat/taajuuslista/vks_5000_frek.html`

7 Discussion

What does a creation like AFE represent and why does it appear to be useful? We pointed out that AFE is linguistically close to the proto-language PBF. Whereas the PBF consists of a normal set of phonemes, AFE has more symbols in it, to the extent that from AFE, one can directly and unambiguously produce Finnish, Estonian and PBF. This extra information proved to be useful when used as a basis for the filtering. The good thing about AFE is that it is simple to produce, and you need not backtrack with your decisions. You simply produce the representation first, and you need not change it. One can automatically produce various versions of the test data: monolingual (FI, ET, AFE), bilingual (FI:ET, FI:AFE, ...). Text versions can be used by the humans and and the FSTs used by the programs.

In spite of the apparent rawness of the AFE representation, it is relatively easy for the linguist to make generalisations using it. A reasonably small set of examples appears to enable the writing quite general constraints. The AFE representation seems to contain most or all the information in the proto-language that is needed for constraining the mappings. The two-level compiler as well as the HFST toolbox and FOMA appear to be extremely useful tools when processing these kinds of language data and relations.

8 Internal reconstruction

A concluding remark is made concerning the two-level framework when representing relations between closely related languages. Suppose that we would have made a morphophonological two-level analyser for Estonian using a similar style as we used in constructing the AFE in this paper. Many Estonian word stems would then have a morphophoneme instead of a vowel at their end. The Estonian word **h a a v** would be represented by its lexical representation such as **h a a v Øa** (because the missing stem final **a** is present in inflected forms) and the Finnish word would be represented as **h a a v ao**. The new kind of AFE would then have a representation like **h a a v ao-aØ**. One can, then, map the Estonian lexical representations more accurately to the AFE representations and therefore, one can map the Estonian words more uniquely into Finnish words.

When one uses morphophonemes in the real languages, one gets even more morphophonemes in this kind of AFE than in the method presented earlier in this paper. Note that using the morphophonemes of individual languages corresponds to the internal reconstruction of the normal historical linguistics. In order to proceed to the phonemes of a realistic proto-language, one establishes the default members of such complex morphophonemes. Thereafter, one can proceed to stepwise establish a proto-language. Filtering the new kind of AFE would be done according to similar methods as above. Similarly the relations between present languages and the proto-language would be constructed according to the principles presented here.

References

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). Openfst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Twelfth International Conference on Implementation and Application of Automata, (CIAA 2007)*, volume 4783 of *Lecture Notes in Computer Science*, pages 11–23, Prague, Czech Republic. Springer.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. Studies in Computational Linguistics, 3. University of Chicago Press. Additional info, see: `www.stanford.edu/~laurik/fsmbook/home.html`.

Bouchard-Côté, A., Griffiths, T. L., and Klein, D. (2009). Improved reconstruction of protolanguage word forms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 65–73, Boulder, Colorado. Association for Computational Linguistics.

Bouchard-Côté, A., Hall, D., Griffiths, T. L., and Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 10.1073/pnas.1204678110.

Campbell, L. (2004). *Historical Linguistics: An Introduction*. Edinburgh University Press, second edition.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece. Association for Computational Linguistics.

Karttunen, L. (1993). Finite-state constraints. In *Proceedings of the International Conference on Current Issues in Computational Linguistics, June 10–14, 1991. Universiti Sains Malaysia, Penang, Malaysia*, pages 173–194.

Kondrak, G. (2002). *Algorithms for Language Reconstruction*. PhD thesis, University of Toronto.

Koskenniemi, K. (1983). *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. Number 11 in Publications. University of Helsinki, Department of General Linguistics.

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T. A., and Silfverberg, M. (2011). Hfst – framework for compiling and applying morphologies. In Mahlow, C. and Piotrowski, M., editors, *Systems and Frameworks for Computational Morphology 2011 (SFCM-2011)*, volume 100 of *Communications in Computer and Information Science*, pages 67–85.

Pirinen, T. (2011). Modularisation of finnish finite-state language description — towards wide collaboration in open source development of a morphological analyser. In Pedersen, B. S., Nešpore, G., and Skadiņa, I., editors, *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011*, NEALT Proceedings Series, Vol. 11 (2011), pages 299–302. Northern European Association for Language Technology (NEALT).

Schmid, H. (2005). A programming language for finite state transducers.

Silfverberg, M. and Lindén, K. (2009). Conflict resolution using weighted rules in hfst-twolc. In *Proceedings of the 17th Nordic Conference of Computational Linguistics, NODALIDA 2009*, pages 174–181. Northern European Association for Language Technology (NEALT).

Wettig, H., Reshetnikov, K., and Yangarber, R. (2012). Using context and phonetic features in models of etymological sound change. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 108–116, Avignon, France. Association for Computational Linguistics.

An SMT Approach to Automatic Annotation of Historical Text

Eva Pettersson[1,2]*, Beáta Megyesi*[1]*, Jörg Tiedemann*[1]

(1) Department of Linguistics and Philology, Uppsala University
(2) Swedish National Graduate School of Language Technology

firstname.lastname@lingfil.uu.se

ABSTRACT

In this paper we propose an approach to tagging and parsing of historical text, using character-based SMT methods for translating the historical spelling to a modern spelling before applying the NLP tools. This way, existing modern taggers and parsers may be used to analyse historical text instead of training new tools specialised in historical language, which might be hard considering the lack of linguistically annotated historical corpora. We show that our approach to spelling normalisation is successful even with small amounts of training data, and that it is generalisable to several languages. For the two languages presented in this paper, the proportion of tokens with a spelling identical to the modern gold standard spelling increases from 64.8% to 83.9%, and from 64.6% to 92.3% respectively, which has a positive impact on subsequent tagging and parsing using modern tools.

KEYWORDS: Digital Humanities, Natural Language Processing, Historical Text, Normalisation, Underresourced Languages, Less-Resource Languages, SMT.

Eva Pettersson, Beáta Megyesi and Jörg Tiedemann 2013. An SMT approach to automatic annotation of historical text. *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013.* NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87: 54–69.

1 Introduction

Historical language may be regarded as an under-resourced language, with limited access to digitized and linguistically annotated corpora, and other NLP resources and tools. Furthermore, it is problematic to develop NLP tools specifically aimed at processing historical text, since the term ”historical” is a wide concept, including texts from a long period of time in which language changes. This means that for example a tagger trained on 14th century text will probably not perform as well on 18th century text. Hence, several taggers would need to be developed for handling different time periods. This problem is further aggravated by the lack of spelling conventions in historical time, meaning that orthography may vary greatly between different authors and genres, or even in the same text written by the same author.

In this paper we propose a method for tagging and parsing of historical text using existing NLP tools developed for modern language. Our approach makes use of standard SMT techniques for normalising the historical spelling to a modern spelling, before applying the NLP tools. Since the machine translation task is performed on a character level, only a small parallel corpus of historical and modern spelling is needed for training.

We show that this method is successful for automatic annotation of historical text, even with small amounts of training data. Furthermore, our two case studies illustrate that this method can be generalised to several languages.

2 SMT-based Tagging and Parsing

The approach to automatic analysis of historical text presented in this paper is illustrated in Figure 1. The input file is a historical text in its original spelling. This text is first tokenised using standard tokenisation. The tokenised text is then normalised to a modern spelling, using character-based SMT, as described further in Section 3. The normalised text is used as input for tagging and parsing, which is performed by available tools for the modern language. In the last step, the annotation is projected back to the original spelling. The final output is thus a tagged and parsed file with the historical spelling preserved.

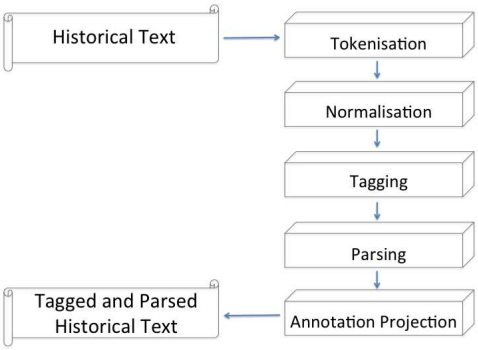

Figure 1: SMT-based tagging and parsing of historical text.

3 Normalisation Using SMT

As stated in Section 1, we regard normalisation of historical text as a translation task. In contrast to traditional translation tasks, normalisation should be performed on a lower level addressing changes in spelling instead of the translation of words and phrases. Hence, treating text normalisation as a case of statistical machine translation leads to a character-level approach without lexical re-ordering. The use of phrase-based SMT in transliteration and character-level translation between closely related languages has already been shown in (Matthews, 2007; Vilar et al., 2007; Nakov and Tiedemann, 2012). We will follow their ideas by applying similar techniques to historical texts.

The basic idea of character-level SMT is that phrases are modeled as character sequences instead of word sequences. Translation models are then trained on character-aligned parallel corpora, and language models on character N-grams. Nakov and Tiedemann (2012) have shown that small parallel training corpora are sufficient for reasonable performance of such an approach. Language models can be trained on larger monolingual corpora, and higher orders in terms of N-gram size can be used to ensure fluent and grammatically correct output (Nakov and Tiedemann, 2012).

The training data in our case is simply a set of word pairs with historical and modern spelling respectively. In contrast to the application of character-level SMT to related language translation, we assume a strict one-to-one correspondence between words of the historical text and its transformation into modern spelling. This simplifies the general setup and ensures that training examples are sufficiently short to enable efficient training procedures.

Important for the success of character-level SMT is the proper alignment of characters that will be used to estimate the parameters of the character-based translation model. We follow the approach of Nakov and Tiedemann (2012) and apply two different tools for this purpose, a weighted finite state transducer implemented in the m2m-aligner (Jiampojamarn et al., 2007), and the word alignment toolkit GIZA++ implementing the IBM models used in statistical MT. Similar to their experiments on related languages, we would like to verify that these alignment techniques are also effective for training spelling normalisation models for historical texts.

The m2m-aligner implements transducer models based on context-independent edit operations. Furthermore, the toolkit allows to include operations over character N-grams (instead of single characters). The transducer is trained using EM on (unaligned) parallel training data. The final model can then be used to produce a Viterbi alignment between given pairs of character strings.

An example is shown in Figure 2, illustrating the alignment of the Icelandic historical spellings *meðr* and *giallda* to the modern versions *meður* and *galda*. In this example, the ϵ symbol denotes empty alignments, i.e. insertions and deletions. For instance, the ϵ symbol in the source word *meðr* denotes the insertion of *u* in the target word *meður*. Likewise, the ϵ symbol in the target word *galda* denotes the deletion of *i* as compared to the source word *giallda*. 2:1 and 1:2 alignments are also possible, as in the case of the alignment of *giallda* to *galda*, where the colon denotes that both letters *l* and *d* in the source word correspond to the single letter *d* in the target word.

```
m|e|ð|ϵ|r|      m|e|ð|u|r|
g|i|a|l|l:d|a|  g|ϵ|a|l|d|a|
```

Figure 2: Character level alignment.

Similarly, GIZA++ is used to train alignment parameters on parallel training data using EM. The toolkit supports several word alignment models which are applied in the order of increasing complexity (see Och (2003) for more details). The final model can then be used again to produce a Viterbi alignment over the data. Word alignment models are certainly not developed for character-level alignment and some of their parameters are not suited for this task. Nevertheless, Nakov and Tiedemann (2012) demonstrate the success of these techniques, which also outperform the transducer-based approach. Therefore, we will revisit these two approaches for our purposes.

Finally, Viterbi alignments are used to extract translations of character sequences and their distributions are used to estimate translation model parameters in the same way as it is done for word-level phrase-based SMT.

4 Experimental Setup

We will run our experiments on two separate data sets: one for Icelandic and one for Swedish, as described further in Section 5 and 6 respectively.

In general, we will focus on phrase-based SMT, i.e. a decomposition of translation models into mappings between non-overlapping character sequences. The SMT engine used is Moses with all its standard components. We apply a phrase-based model with the common feature functions: 2 phrase translation probabilities, 2 lexical weights, a phrase penalty feature, a language model feature and a word penalty. The feature weights are trained using MERT with BLEU over character-sequences as the objective function. The phrase table scores are produced using the Moses phrase extraction and scoring methods. The maximum size of a phrase (sequence of characters) is set to 10. Language models (10-gram models) are estimated using SRILM (interpolated and smoothed) which are then transformed into binary versions to be used by KenLM during decoding. Reordering is switched off during decoding (and tuning) as we assume monotonic alignments.

Training character-based SMT models for text normalisation basically involves the following steps (see further Section 3):

- **Word Alignment**
  Aligning corresponding words that will serve as the parallel training data for creating translation models.
- **Character alignment**
  Aligning characters for the entire corpus of aligned word pairs.
- **Language modeling**
  Training character-based language models on monolingual data in the target language.
- **Parameter tuning**
  Tuning character-based SMT on some development data.

Example: *dömmes (historical) – döms (modern)*

aligned bigrams:

```
dö öm mm me es s_
 |  |  \/   | /
dö öm  ms   s_
```

mapped to character alignments:

```
d ö m m e s
| | |/  |/
d ö m   s
```

Figure 3: Character-bigram alignment between a Swedish word in its historical and modern spelling respectively.

There are various settings and approaches that can be tested for each of these steps. In our experiments we will look at some important aspects of performing these tasks.

Interestingly, it is possible to adapt well-known alignment techniques to a lower level when moving from word-based to character-based models as we will see in our experiments. Word alignment can in fact be modeled in the same way as sentence alignment is modeled otherwise and character alignment can be performed successfully using word alignment techniques.

For word alignment, we can assume a strong correlation between the length of the original spelling and the modern spelling of the same word. Futhermore, we can assume that corresponding words contain many corresponding characters as well and that there should be no re-ordering of words in the normalised version, which leads to monotonic alignments. This is exactly what is used in common length-based and lexical matching-based sentence alignment algorithms. We therefore apply a standard sentence alignment tool for the word alignment task, *hunalign* (Varga et al., 2005), that combines both features; length correlation and lexical matches.

Character alignment is a bit more tricky and the relation to word alignment is less clear. In text normalisation, we would definitely not expect a lot of distortion (which is an important part of common word alignment models) and the vocabulary size is not comparable to that of a word alignment task as pointed out by Nakov and Tiedemann (2012). We therefore include the extension to bigrams as suggested by the same authors and also compare the results with transducer-based alignments that are developed for character-level transformations.

Character-bigrams include a bit more contextual information that may help the alignment, which otherwise mainly uses context-independent parameters. Fortunately, links between bigrams can still easily be mapped back to single character alignments which we need for training our translation models. See Figure 3 for an illustration of this process.

Similar to standard phrase-based SMT, we then extract mappings between character sequences of up to seven characters which are consistent with the alignment between characters. The standard scoring techniques are applied as in word-based models.

Language modeling can be done in the same way as for word-based models but trained on character sequences. We use 10-grams as suggested by Nakov and Tiedemann (2012) and standard smoothing techniques.

Parameter tuning is the final task that needs to be performed for building the character-based SMT model. We experiment with a standard Minimum Error Rate Training (MERT) optimized towards BLEU over character sequences on the translation of single words. We also tried sentence-based decoding with word-level BLEU but the normalisation accuracy was lower than with our word-based decoding.

4.1 Evaluation

The performance of our SMT-based approach to tagging and parsing of historical text is evaluated in three aspects: 1) normalisation accuracy, i.e. the percentage of tokens in the automatically normalised version of the text that are identical to the manually modernised gold standard version, 2) tagging performance before and after normalisation, and 3) parsing performance before and after normalisation.

We will look at different alignment settings and data sizes when evaluating the normalisaton accuracy. Furthermore, we will compare the normalisation accuracy to two baselines, hereafter referred to as *unnormalised* and *baseline*:

1. **unnormalised**
   The proportion of tokens in the unnormalised source text with a spelling identical to the modern spelling.
2. **baseline**
   The normalisation accuracy achieved when simply replacing each historical word type with its most frequent modern version observed in the word-aligned training corpus (leaving previously unseen tokens unchanged).

First, we will look at the case of historical Icelandic texts before moving on to a case study on Swedish texts. Since the access to corpora and NLP tools differs between the languages, the experiments are not entirely the same for the two languages, and the results are thus not fully comparable between the languages. However, for both languages we show that the proposed approach is successful for automatic annotation of historical text.

5 Case Study 1: Icelandic

For training and tuning of the Icelandic character-based SMT system, we make use of Snorri Sturluson's *Edda* (the Uppsala version DG11), an Icelandic saga available both in its 14th century spelling and in a manually modernised version (Palsson, 2012). These versions were automatically aligned using hunalign, and the extracted alignments were manually corrected, resulting in a parallel corpus of 33,888 entries in total. Training and tuning sets were created by extracting every 10th sentence to the tuning set, and the rest of the sentences to the training set. This extraction resulted in a training set of 30,451 token pairs and a tuning set of 3,437 token pairs.

As a language model for the SMT system, we use a subset of the Tagged Icelandic Corpus of contemporary Icelandic texts (Helgadóttir et al., 2012). In total, this subset consists of 21,613,551 tokens distributed over 12 genres, including for example newspaper text, blog text,

parliamentary speeches and university essays. We would of course prefer to use the whole corpus, but at the time of writing, the full corpus has not yet been made available.

Evaluation of normalisation accuracy, as well as tagging results, is performed on a subset of *Ectors saga* from the 15th century (Loth, 1962). This text contains 20,811 tokens and is part of the Icelandic Parsed Historical Corpus, IcePaHC, a manually tagged and parsed diachronic corpus of texts spanning from 1150 to 2008 (Rögnvaldsson et al., 2012). In the IcePaHC corpus, the old texts have been manually modernised with regard to spelling. We also have access to the saga in its 15th century spelling, needed for evaluation purposes.

For tagging, we use the IceNLP tagger (Loftsson and Rögnvaldsson, 2007) trained on the Icelandic Frequency Dictionary corpus (IFD) of approximately 500,000 tokens from the time period 1980–1990 (Pind, 1991).

Evaluation of parsing performance was not possible, since we do not have access to a parsed gold standard corpus in a suitable format. The IcePaHC corpus does include syntactic information, but in an annotation scheme that is not easily mapped to the one produced by the IceNLP parser.

5.1 Word Alignment

To train a character-based SMT system on the *Edda*, the historical and the modern version of the text has to be aligned on a word level. For this purpose we use hunalign (Varga et al., 2005), which is in fact a sentence aligner rather than a word aligner. As stated in section 4, for our specific alignment task we can however assume monotonic alignments without re-ordering and with a strong correlation between the length of the original spelling and the modern spelling, which would be a suitable task for sentence alignment. We tried four different ways of using hunalign for our word alignment task:

1. **no split**
   Input data is one token on each line, with empty lines denoting sentence boundaries.
2. **no split +realign**
   Same as ”no split”, but with the additional flag *-realign*, in which the aligner is run in three phases, heuristically building a dictionary based on the identified sentence pairs (in our case word pairs).
3. **split**
   Same as ”no split”, but with whitespace separating all characters.
4. **split +realign**
   Same as ”split”, but with the additional *-realign* flag, as described above.

Automatic alignment will inevitably introduce noise in the training data. Since the modern version is stated to be a modernisation of spelling, not including syntactic normalisation, the word alignment task is however easier than in an ordinary translation setting. To evaluate the impact of the different alignment methods on the end result, we ran experiments based on the GIZA++ unigram setting for character alignment, with training data automatically generated from the four alignment methods described above, as well as with training data that had been manually corrected. As illustrated in Table 1, splitting the words into their separate characters has a positive effect on the alignment results. With the split +realign setting, normalisation accuracy is 79.2% as compared to 78.8% for the no split setting.

Approximately two thirds (64.8%) of the original tokens already had a spelling that was identical to the modern spelling. With the simplistic baseline, replacing each historical word type with its most frequent modern version observed in the word-aligned training data, normalisation accuracy increases to 69.8%, which is far from the results achieved for the SMT model.

It is also worth mentioning, that the normalisation accuracy for automatically generated training data is close to the accuracy achieved for manually corrected training data; 79.2% in the best hunalign setting as compared to 83.9% for manually corrected data. Hence, if no word-aligned historical-modern data is available, automatic alignment techniques may successfully be used to automatically create such a parallel corpus.

Hunalign Model	**Accuracy**
unnormalised	64.8%
baseline	69.8%
no split	78.8%
no split +realign	78.8%
split	79.1%
split +realign	79.2%
manual alignment	83.9%

Table 1: Normalisation accuracy with different methods for alignment of the training data. Unnormalised = Percentage of tokens in the unnormalised text with a spelling identical to the modern spelling. Baseline = Normalisation accuracy when replacing each historical word type with its most frequent modern version observed in the word-aligned training corpus.

5.2 Character Alignment

As mentioned in Section 3, we have experimented on using GIZA++ and m2m aligner models with different settings for training the character-based SMT systems. The results for different settings on the Icelandic training corpus are summarised in Table 2.

Alignment	**Accuracy**
unnormalised	64.8%
baseline	69.8%
giza unigram	83.9%
giza bigram	83.5%
m2m 1:1	81.0%
m2m 2:2	83.6%

Table 2: Normalisation accuracy for different character alignment models. Unnormalised = Percentage of tokens with a spelling identical to the modern spelling before normalisation. Baseline = Normalisation accuracy when replacing each historical word type with its most frequent modern version observed in the word-aligned training corpus. Giza uses standard word alignment models (for character unigrams and bigrams) and m2m uses the WFST with single character edit operations (1:1) and multi-character operations (2:2).

As previously stated, approximately two thirds (64.8%) of the original tokens already had a spelling that was identical to the modern spelling. One could have expected this figure to be higher, since modern Icelandic is generally seen as close in spelling to its historical variants. However, in addition to spelling variation, Icelandic has a rich morphology that has changed over time, which is also influencing the words in terms of string similarity. The SMT models were able to successfully capture some of these differences, resulting in 83.9% correctly normalised tokens in the best setting, i.e. the GIZA++ unigram setting. Using more informative bigrams instead of unigrams surprisingly lead to a drop in normalisation accuracy from 83.9% to 83.5%. We also tried the m2m aligner, with single character edit operations (1:1) and multi-character operations (2:2), which was slightly less successful for our normalisation task as compared to the GIZA++ unigram model.

5.3 Training Data

In Section 5.2, we have shown that training a character-based SMT system on a parallel corpus of historical and modern spelling is successful for normalising historical text to a modern spelling. So far, we have used a parallel corpus of 33,888 token pairs for training and tuning. It might be the case that only a smaller data set is available in this form, or that such a parallel corpus is not available at all and would need to be manually created. To see whether an even smaller corpus would be sufficient for achieving reasonable results, we experimented on different sizes of the parallel corpus used for training. As expected, the more training data, the better normalisation results (in general), as presented in Figure 4. However, with only 1,000 token pairs, we already achieve 76.5% normalisation accuracy, as compared to 83.9% for the entire corpus.

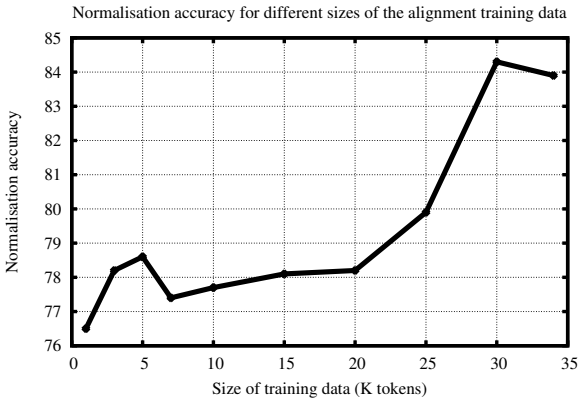

Figure 4: Normalisation accuracy when varying the size of the alignment training data.

5.4 Language Modeling

For the language model, we have experimented on varying the size and genres included in the training data. A motivation for this is that we would like our normalisation approach to be useful for any language where the modern version of the language has a Basic Language Resource Kit, BLARK (as defined in Krauwer et al. (2004)). Such a BLARK may contain corpora of varying sizes, sometimes including several genres and sometimes including for example newspaper text only. We have performed experiments on varying the size of the language model training data for Icelandic, from 1 million tokens as a minimum to including all 21,613,551 tokens of the corpus. Furthermore, we have experimented on using newspaper text only, as compared to including all the genres of the corpus. The results are presented in Figure 5.

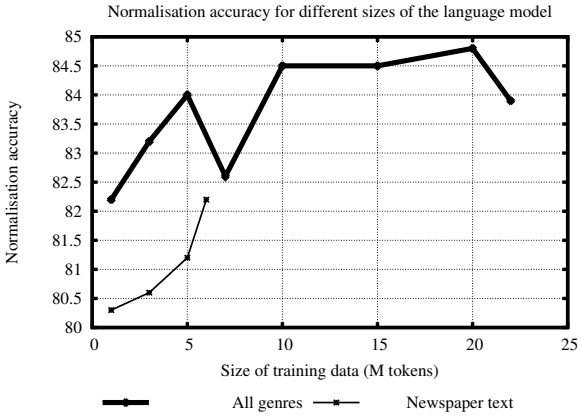

Figure 5: Normalisation accuracy, when varying the size and text types included in the language model.

As seen from the results, including a corpus of different genres in the language model shows slightly better results than using newspaper text only. However, there is not a huge difference between the two types of corpora. For 5 million words of newspaper text, a normalisation accuracy of 81.2% is achieved, as compared to 84.0% for the sampled corpus. It is also worth mentioning that whereas the newspaper corpus shows the expected increase in normalisation accuracy when more data is added, this relation is not as clear-cut for the sampled corpus where the addition of more data in some cases leads to a drop in normalisation accuracy. This could be due to larger variation in the sampled training data, meaning that adding new data sometimes distort the language model.

Also note that including only 1 million (sampled) tokens results in a normalisation accuracy of 82.2%, which is already close to the 83.9% achieved when including the full corpus in the language model.

5.5 Tagging

Without normalisation, the IceNLP tagger is able to correctly annotate 46.7% of the tokens in the test text (*Ectors saga*). Besides the spelling variation, one reason for the rather low tagging

accuracy could be that the manually tagged corpus is, as stated earlier, based on modernised spelling. The modern spelling also seem to include morphological changes, meaning that the gold standard tag is in fact not in all cases the correct tag for the token when its historical spelling is preserved. Despite this deficiency in the gold standard corpus, tagging accuracy increases by 10 percentage units after normalisation, to 56.6%, see further Table 3.

	Accuracy
unnormalised	46.7%
baseline	49.9%
normalised	56.6%

Table 3: Accuracy for Icelandic part-of-speech tagging. Baseline = Normalisation by replacing each historical word type with its most frequent modern version observed in the word-aligned training corpus. Normalised = Normalisation by the Giza unigram approach.

6 Case Study 2: Swedish

For training, tuning and evaluation in the Swedish setting, we use a corpus of court records and church documents from the period 1527–1812. These texts are available in both their original and their modernised spelling (see further Pettersson et al. (2012) for more details on the corpus). For training we extracted 28,237 token pairs, whereas the tuning set consists of 2,590 (non-overlapping) token pairs and the test set includes 33,544 (non-overlapping) token pairs, equally distributed over the texts in the corpus. The test set is the same subset of the corpus as is used in Pettersson et al. (2012).

As a language model, we make use of the Stockholm-Umeå Corpus, SUC, a balanced corpus consisting of approximately one million tokens extracted from a number of different text types representative of the Swedish language in the 1990s (Ejerhed and Källgren, 1997).

For tagging, we use HunPOS (Halácsy et al., 2007), a free and open source reimplementation of the HMM-based TnT-tagger by Brants (2000). In our experiments, we use HunPOS with a Swedish model based on the SUC corpus.

For parsing, we use MaltParser version 1.6, a data-driven dependency parser developed by Nivre et al. (2006a). In our experiments, the parser is run with a pre-trained model[1] for parsing contemporary Swedish text, based on the Talbanken section of the Swedish Treebank (Nivre et al., 2006b).

6.1 Character Alignment

We have experimented on using GIZA++ and m2m aligner models with different settings for training the character-based Swedish SMT systems, as presented in Table 4. Similar to the results achieved for Icelandic, the best-performing setting is to train a GIZA++ unigram model, which outperforms the more informative bigram model as well as the m2m aligner models. In the GIZA++ unigram setting, the percentage of tokens with a spelling identical to the gold standard spelling increases from 64.6% for the original spelling, to 92.3% after normalisation. It is interesting to note that the baseline method of simply replacing each historical word type with its most frequent modern version observed in the word-aligned training corpus also

[1] http://maltparser.org/mco/swedish_parser/swemalt.html

has a large positive effect on normalisation accuracy, increasing the number of tokens with a modern spelling from 64.6% to 86.1%. For Icelandic this effect was considerably less significant, increasing the number of tokens with a modern spelling from 64.8% to 69.8%. One reason for the larger impact on Swedish might be that training and test sets are extracted from the same historical corpus (though non-overlapping), whereas for Icelandic the training set is extracted from one text and the test set from another text.

Alignment	**Accuracy**
unnormalised	64.6%
baseline	86.1%
giza unigram	92.3%
giza bigram	91.8%
m2m 1:1	91.4%
m2m 2:2	89.7%

Table 4: Normalisation accuracy for different character alignment models. Unnormalised = Percentage of tokens with a spelling identical to the modern spelling before normalisation. Baseline = Normalisation accuracy when replacing each historical word type with its most frequent modern version observed in the word-aligned training corpus. Giza uses standard word alignment models (for character unigrams and bigrams), and m2m uses the WFST with single character edit operations (1:1) and multi-character operations (2:2).

6.2 Tagging and Parsing

For the Swedish setting, we do not have access to a linguistically annotated gold standard for evaluating tagging and parsing performance. However, in the test corpus described in Section 6, all the verbs and their complements have been manually annotated as such. Therefore, we decided to indirectly evaluate the performance of a tagger based on precision and recall for verb identification before and after normalisation of the input text. Similarly, parsing performance is indirectly evaluated based on the proportion of correctly detected verb complements.

Tagging The results presented in Table 5 show that normalisation has a large positive impact on precision and recall measures for verb identification. Without normalisation, an F-score of 70.4% is achieved for this task, as compared to 83.5% for the baseline approach to normalisation and 88.7% for the SMT approach to normalisation. The largest increase is in recall, where 89.8% of the verbs are identified after normalisation, as compared to 64.2% for the original spelling. Precision also increases substantially, from 77.9% to 87.7%.

Parsing MaltParser produces dependency trees labeled with grammatical functions, which can be used to identify different types of complements. In this setup, we define *exact matches* as cases where the detected verb complement is identical to the corresponding gold standard complement. Likewise, we define *partial matches* as cases where the detected verb complement has a non-empty overlap with the corresponding gold standard complement. Consider for example the gold standard analysis *effterfrågat [om sinss manss dödh]* (asked [about her husband's death]). A system output where the head noun of the complement is missing, will still be regarded as partially correct: *effterfrågat [om sinss manss]* (asked [about her husband's]).

	Precision	Recall	F-score
unnormalised	77.9%	64.2%	70.4%
baseline	84.0%	83.0%	83.5%
normalised	87.7%	89.8%	88.7%

Table 5: Precision and recall measures for verb identification based on tagging. Baseline = Normalisation by replacing each historical word type with its most frequent modern version observed in the word-aligned training corpus. Normalised = Normalisation by the Giza unigram approach.

As shown in Table 6, the proportion of correctly identified verb complements (exact and partial matches) increases from 32.9% for the original historical spelling to 42.6% after baseline normalisation, and 46.2% after SMT normalisation.

	Exact Match	Partial Match	Exact or Partial Match
unnormalised	23.8%	9.1%	32.9%
baseline	30.2%	12.4%	42.6%
normalised	33.2%	13.0%	46.2%

Table 6: Precision and recall measures for verb complement detection based on parsing. Baseline = Normalisation by replacing each historical word type with its most frequent modern version observed in the word-aligned training corpus. Normalised = Normalisation by the Giza unigram approach.

7 Previous Studies

There are several approaches to spelling modernisation of historical text that have been described in previous studies. One of the earlier examples is the VARD tool (VARiant Detector), which is based on dictionary lookup, mapping 16th to 19th century English spelling to modern spelling. Evaluation was performed on a set of 17th century texts, and compared to the performance of modern spell checkers. Between a third and a half of all tokens were correctly normalised by both VARD and MS Word, whereas approximately one third of the tokens were correctly normalised only when using VARD. The comparison between VARD and Aspell showed similar results (Rayson et al., 2005).

Pettersson et al. (2012) tried a rule-based letter transformation approach, where a set of 29 hand-crafted normalisation rules was produced on the basis of a 17th century court records text. The resulting rule set was applied to a gold standard corpus of 33,544 tokens from the period 1527–1812, i.e. the same test corpus as is used in our Swedish case study. The normalisation had a positive effect on texts from all centuries covered in the corpus. On average, approximately 73% of the tokens were correctly normalised using this method.

A data-driven approach based on Levenshtein similarity was presented by (Bollmann et al., 2011) for normalisation of Early New High German. Normalisation rules were automatically derived by means of the Levenshtein edit distance, based on a word-aligned parallel corpus consisting of the Martin Luther bible in its 1545 edition and its 1892 version, respectively. Using this normalisation technique, the proportion of words with a spelling identical to the modern spelling increased from 65% in the original text to 91% in the normalised text.

8 Conclusion

In this paper, we have shown that using character-based SMT techniques for normalising historical text to a modern spelling, is successful when applying modern taggers and parsers to analyse historical text. In the Swedish case study, the proportion of tokens in the historical text that are identical to the modern spelling increases from 64.6% to 92.3% in the best setting. This results in an increase in F-score for verb identification (based on tagging) from 70.4% before normalisation to 88.7% after normalisation. Accordingly, the proportion of correctly identified verb complements (based on parsing) increases to the same extent.

Furthermore, we have shown that it is possible to achieve good results without (or with little) manual efforts, since only a small amount of training data is needed to achieve reasonable results. With only 1,000 tokens available in a historical and a modern spelling, a normalisation accuracy of 76.5% is achieved for Icelandic, as compared to 83.9% in the best setting. If no word-aligned training data at all is available, automatic sentence alignment methods may successfully be used for automatically creating training data, as shown in section 5.1.

The data-driven nature of our approach makes it language-independent, and we believe that our method is generally applicable to languages for which there is a corpus of modern text available, as well as a small data set of historical texts with both historical and modern spelling. So far, we have evaluated our approach for Swedish and Icelandic. In the future, we would like to extend our experiments to texts from other languages, time periods and genres. It would also be interesting to explore ways of normalising not only spelling, but also morphological and syntactic differences in the historical texts.

References

Bollmann, M., Petran, F., and Dipper, S. (2011). Rule-based normalization of historical texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42, Hissar, Bulgaria.

Brants, T. (2000). TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)*, Seattle, Washington, USA.

Ejerhed, E. and Källgren, G. (1997). Stockholm Umeå Corpus. Version 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. ISBN 91-7191-348-3.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 209–212, Prague, Czech Republic.

Helgadóttir, S., Svavarsdóttir, A., Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. (2012). The tagged icelandic corpus (mím). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages*, pages 67–72.

Jiampojamarn, S., Kondrak, G., and Sherif, T. (2007). Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, pages 372–379, Rochester, NY.

Krauwer, S., Maegaard, B., Khalid, C., and Damsgaard Jørgensen, L. (2004). Report on Basic Language Resource Kit (BLARK) for Arabic.

Loftsson, H. and Rögnvaldsson, E. (2007). IceNLP: A natural language processing toolkit for Icelandic. In *Proceedings of InterSpeech, Special session: Speech and language technology for less-resourced languages*, Antwerp, Belgium.

Loth, A., editor (1962). *Late Medieval Icelandic Romances I*. Kaupmannahöfn, Copenhagen.

Matthews, D. (2007). Machine transliteration of proper names. Master's thesis, School of Informatics.

Nakov, P. and Tiedemann, J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–305, Jeju Island, Korea. Association for Computational Linguistics.

Nivre, J., Hall, J., and Nilsson, J. (2006a). MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC)*, pages 2216–2219, Genoa, Italy.

Nivre, J., Nilsson, J., and Hall, J. (2006b). Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC)*, pages 24–26, Genoa, Italy.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of ACL'03*, pages 160–167, Sapporo, Japan.

Palsson, H., editor (2012). *The Uppsala Edda*. Viking Society for Northern Research.

Pettersson, E., Megyesi, B., and Nivre, J. (2012). Rule-based normalisation of historical text - a diachronic study. In *Proceedings of the First International Workshop on Language Technology for Historical Text(s)*, Vienna, Austria.

Pind, J., editor (1991). *Icelandic Frequency Dictionary*. Institute of Lexicography, Reykjavik, Iceland.

Rayson, P., Archer, D., and Nicholas, S. (2005). VARD versus Word – A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. In *Proceedings from the Corpus Linguistics Conference Series on-line e-journal*, volume 1, Birmingham, UK.

Rögnvaldsson, E., Ingason, A. K., sson, E. F. S., and Wallenberg, J. (2012). The icelandic parsed historical corpus (icepahc). In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP*, pages 590–596.

Vilar, D., Peter, J.-T., and Hermann, N. (2007). Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague, Czech Republic. Association for Computational Linguistics.

Edit Transducers for Spelling Variation in Old Spanish

Jordi Porta, José-Luis Sancho, Javier Gómez

Departamento de Tecnología y Sistemas
Centro de Estudios de la Real Academia Española
c/ Serrano 187-189, Madrid 28002. Spain

{porta,sancho,javierg}@rae.es

ABSTRACT
A system for the analysis of Old Spanish word forms using weighted finite-state transducers is presented. The system uses previously existing resources such as a modern lexicon, a phonological transcriber and a set of rules implementing the evolution of Spanish from the Middle Ages. The results obtained in all datasets show significant improvements, both in accuracy and in the trade-off between precision and recall, with respect to the baseline and the Levenshtein edit distance. A qualitative error analysis suggests several potential ways to improve the performance of the system.

KEYWORDS: Old Spanish, Finite-State Transducers, Spelling Variation, Historical Linguistics.

Jordi Porta, José-Luis Sancho and Javier Gómez 2013. Edit transducers for spelling variation in Old Spanish. *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013.* NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87: 70–79.

1 Introduction

When processing historical language variants, the most visible problem is the lack of a standardised orthography. The spelling of texts written in different periods of time varies because spelling conventions change over time and official orthographies, when exist, are periodically subjected to reforms. In addition, texts written during the same period of time also have been found to have variation in spelling. And, as if that were not enough, this variation can be found within the same text and even within works by the same author.

In this work, we address the problem of assigning modern citation forms (or lemmas) and word classes to historical word forms using a system for the treatment of diachronic variation found in Old Spanish. As it has been pointed out in Borin and Forsberg (2008), the assignment of word forms to citation forms is seen as a morphological analysis differing from part-of-speech tagging. While the former provides all the plausible analyses for a given word, the later assigns to words the most probable morphological analysis given the context.

The system is modular and has been implemented with weighted finite-state transducers and, in its current state, contains the devices for dealing with the phonological evolution and orthographic variation of Old Spanish. It uses also a Contemporary Spanish lexicon providing the analysis. Several experiments with different datasets have been conducted in order to assess the validity of the approach and results have been compared with those obtained using only the lexicon and in combination with the Levenshtein distance.

2 A Very Brief History of the Spanish Language

Traditionally, different periods of time are distinguished in the history of a language. In the case of the Spanish language these are: Pre-literary Romance, Medieval Spanish, Golden Age, etc. However, neither the periodisation nor the so-called 'language stages' are justified in the light of internal factors. Linguistic change may take centuries to complete and can take place in different communities at different times. The progressive abandonment of Latin in favour of the Romance languages took place during the Middle Ages, but Latin was still used as a lingua franca and in literate culture for long time. A more complete, detailed and adequate history of the Spanish language can be found in Penny (2002).

This paper focuses on historical words found in texts dated in Medieval and Golden Ages. One of the most important morphosyntactic changes of Romance from Latin is the major reduction in the nominal case system. At the syntactic level, the functions of declension were taken on by a system of prepositions and other particles. Another change is the development of synthetic future and conditional paradigms from analytic constructions in which mesoclisis becomes proclisis. According to Lloyd (1987), the development of the Old Spanish phonological system from Late Latin to Contemporary Spanish is best described as the interplay of many relaxing and simplifying processes: some vowel distinctions were dropped resulting in the current five vowel system while other vowels simply disappeared as a result of syncope, elision or apocope (notably final *-e*), to name just a few. On the consonant wagon, the relaxing trend is responsible for consonant cluster simplification, debuccalisation (consonant resulting in a vowel), palatalisation, lost of initial *f-* or various kinds of neutralisation (notably *r*/*l*). The so-called lenition, really a cascaded set of changes from geminate simplification, plosive spirantisation and voicing, also progressed to easier articulation. The most outstanding change is, by far, the devoicing of the three sibilant series that finally led to the current system. The orthographic system reflected the sound changes with a certain lag, making writing practices confusing. According to Morreale

(1978), the configuration of 'medieval spelling' is the result of the interplay of palaeographic usages (the letter shapes), graphical usages (the letter identification) and phonetic values. Experts have identified three principles governing the correspondence between phonological units and the spelling representing them: pronunciation, etymology and usage (Pombo, 2012). Although Spanish favours phoneticism from as early as 13th century, texts show alternating spelling trends rooting on dialectal, cultural (i.e. etymological trend in Late Middle Ages) or stylistic (i.e. *variatio*) reasons. As a result, we are faced with what Pombo (2012) calls *heterographs* (several spellings for one sound) and *heterphones* (several sounds for one spelling).

3 Previous Related Approaches to Spelling Variation

A recent survey on the problems and the approaches to the processing of historical texts, from their acquisition to their exploitation, can be found in Piotrowski (2012). The task of analysing historical word forms using modern lexicons has been approached as an approximate string matching problem. In approximate matching an edit distance is used to measure the similarity between two strings and is defined as the minimal cost required to transform one string into another. The most commonly used distance is known as the Damerau-Levenshtein distance. The basic editing operations considered in the Levenshtein distance are the deletion, the insertion and the substitution of a letter (Levenshtein, 1966). All these operations are assigned a unit cost. Some other operations are often considered as basic edit operations, e.g. letter transposition (Damerau, 1964), which are useful correcting errors made during the fast keyboarding of texts.

In Bollmann et al. (2011), character replacement rules were derived from the alignment of the Luther's 1545 bible translation in Early New High German and a corresponding version of the bible in New High German. Normalisation for Luther texts results in 91% exact token matches (and 93% adding a word substitution list). However, performance went down when the time period was extended and the number of different authors considered was increased. For different versions of the *Interrogatio Sancti Anselmi de Passione Domini*, written between the 14th and 16th centuries, the number of exact matches at token level was of about 42%, but started in a baseline of 32%.

The approach presented here has many points in contact with Jurish (2010b). In his work, a historical word form is canonised by a modern word if they conflate through the application of a cascade of transducers implementing transliteration, phoneticization and heuristic rewrites. In order to increase precision, the context of a historical word is taken into account to disambiguate conflated analysis. For a 1.5 million word corpus of 18–19th century German, precision and recall reached at token level were of 94.3 and 99.3% respectively.

4 A Linguistic Approach Based on Transducers

Analysing a word in our model means computing the result of the composition of three transducers:

$$W \circ E \circ L \tag{1}$$

where W is the automaton representing the word, E is in general any edit transducer, and L is a lexicon relating word forms and its correspondent lemma and word class. E is currently used to model phonological and graphical variation. Note that if W is a historic word form and L is composed of modern forms, there is an implicit process of canonicalisation that leads to the modern analysis. The word analyser is implemented by merging an on-demand version of the three-way composition algorithm (Allauzen and Mohri, 2008) with the n-best strings algorithm of Mohri and Riley (2002) adapted to return only the strings with lowest cost. A very similar

solution has been also proposed in Jurish (2010a). The system has been implemented using the OpenFst library (Allauzen et al., 2007).

The *Diccionario de la lengua española de la Real Academia Española*[1] (DRAE (RAE, 2001)) is a general lexical repertoire focused on the elevated norm shared by all educated speakers all over the Spanish speaking areas. It is an ever evolving, periodically updated dictionary, originated in the early 18th century. This longevity, together with the lack of a historical dictionary that fully documents lexical and semantic evolution, account for the density of both lemmas included and grammatical and semantic distinctions made: long term unused words or meanings, words or senses specific only to some countries or areas, mildly used technical terms or even latin formulae are given an entry. On the other side, regular, on occasion very widespread words resulting from morphological processes are missing. While this impairs the reusability of the dictionary on natural language processing tasks, it is the reference dictionary of choice and as such we decided for most of the experiments to keep full contents activated, i.e. words were not selectively deactivated by chronological or geographical criteria. The lexicon derived from the 2012 amended version of DRAE has been augmented with the forms resulting from certain morphological processes: *-mente* (-ly) adverbs, superlative and augmentative or diminutive derivatives. Unlike Contemporary Spanish, Old Spanish allowed clitic affixation to any verbal form. In order to account for those complexes, we generated about 50 million entries from the combination of every verbal form with up to three enclitic pronouns.

The edit transducer plays the role of variation model. We have implemented two edit transducers for the purpose of comparison: the Levenshtein transducer and a rule-based linguistic transducer implementing the sound change in Spanish. An extensive use of rewrite rules of the kind of $\alpha \rightarrow \beta \ / \ \gamma \ \_\_\_ \ \delta$ (Chomsky and Halle, 1968) is done in the descriptions of language at different levels. Rewrite rules turn out to be efficiently compiled into finite-state transducers (Kaplan and Kay, 1994; Karttunen, 1995, 1996). The linguistic transducer implements a cascade of applications of rewrite rules organised in modules. The overall process can be summarised as follows: First, modern word forms in the lexicon are phonetically transcribed, then a set of rules expressing the phonetic and phonological change is applied and the resulting forms are transcribed back to graphemes. Finally, a set of rules for graphical variation is applied. The application of these modules can be formulated in terms of regular relations as:

$$E = (M \circ P \circ C \circ G \circ V)^{-1} \tag{2}$$

where

- M is a set of modern word forms.
- P represents a grapheme-to-phoneme transcription containing rules as the following, which map the letter c to the SAMPA (Wells, 1997) symbols $/T/$ or $/k/$ depending on the context:
$$(c \rightarrow /T/ / \_\_\_ \{E, I\}) \circ (c \rightarrow /k/) \tag{3}$$
- C are the set of rules expressing phonological change. For example, the series $/ds/ > /Z/ > /S/ > /T/$ and $/ts/ > /S/ > /T/$ accounts for the evolution of palatal affricates,

[1] `http://lema.rae.es/drae`

where deaffrication, devoicing (when applicable) and dentalisation are at play. Their implementation is as follows:

$$(/ds/ \rightarrow /Z/) \circ (/ts/ \rightarrow /S/) \circ (/Z/ \rightarrow /S/) \circ (/S/ \rightarrow /T/) \quad (4)$$

- G translates phonological forms into surface forms. The following example maps $[k]$ to its alternating graphemic realisations:

$$(/k/ \rightarrow qu/\_\_\{E,I\}) \circ (/k/ \rightarrow c) \quad (5)$$

- V contains graphemic equivalences, as in the following example, where $(\rightarrow)$ makes the rewriting optional and (0.2) is the weight associated to the rewrite rule:

$$(\{c,z\} \rightarrow \{s,z\}/\_\_\{E,I\}) \circ (z\ (\rightarrow)\ ç\ (0.2)/\_\_\{A,O,U\}) \quad (6)$$

It is worth noting that G differs from P^{-1} in that G implements a relaxed form of the current orthographic norms. Note also that in order to get a transducer from Old Spanish to Modern Spanish the result of the composition is inverted $(^{-1})$. These weighted rational relations are all expressed using regular expressions and context-dependent rewrite rules. Preference in rewriting is expressed using numerical values or weights in the tropical semiring (Mohri, 2009). All these rules and regular expressions are compiled into weighted finite-state transducers. For its implementation we have used the OpenGrm Thrax Grammar Compiler (Roark et al., 2012).

5 Datasets

Experiments have been conducted on different datasets of Medieval and Golden Ages that represent different gold standards. We want to note that these datasets are word lists and that we report figures on a type basis, unlike other works reporting results on running texts, i.e. on a token basis. The distribution of word classes of these datasets can be seen in Table 1.

The dataset called FL-EM basically corresponds to the lexicon found within the FreeLing[2] distribution for analysing Old Spanish. The creation of this lexicon containing variants observed in the corpus of medieval texts of the Hispanic Seminary of Medieval Studies is described in Sánchez-Marco et al. (2011). It is important to note that FreeLing bases the analysis of some words on morphological decomposition modules: verbs with enclitics, adverbs ended in *-mente*, augmentatives and diminutives, superlatives, etc. Therefore, these words are not found in the FreeLing static lexicon. Proper nouns and Roman numerals, as well as multiword and amalgamated expressions have been removed for the experiments. Neither the system of words classes nor the lemmatisation is directly comparable and some categories and lemmas of the DRAE have been changed before comparisons not to get false differences in the experiments. Notably, old lemmas from DRAE with modern counterparts were deactivaded, e.g. *fermosura*, whose modern lemma is *hermosura* (beauty).

CDH-EM and CDH-SO are the lexicons induced from the manual correction of a previous annotation of a subcorpus of the Spanish corpus *Corpus del Nuevo diccionario histórico*[3] (CDH). CDH-EM comes from a fragment of 67 661 running words from medieval texts spanning from 1064 to 1494 and CDH-SO contains 318 728 running words from Golden Ages texts (1521–1698).

[2] `http://nlp.lsi.upc.edu/freeling/` (accessed 2013-03-03)

[3] `http://www.frl.es/Paginas/Corpusdiccionariohistorico.aspx` (accessed 2013-03-03)

Word Class	FL-EM
Adjectives	4 048
Nouns	11 257
Verbs	20 339
Prepositions	64
Determinants	172
Pronouns	292
Adverbs	254
Conjunctions	160
Interjections	117
Other	6
Total	36 709

Word Class	CDH-EM	CDH-SO	MAP-EM	MAP-SO
Adjectives	1 714	4 974	10 230	5 218
Nouns	3 505	9 855	23 776	11 533
Verbs	5 967	16 046	45 021	22 275
Prepositions	55	56	153	83
Articles	12	15	24	22
Pronouns	235	319	839	353
Adverbs	274	476	1 459	555
Conjunctions	52	52	198	103
Interjections	114	169	310	145
Other	1	6	13	10
Total	11 929	82 023	31 968	40 297

Table 1: Word class distributions in the FL-EM dataset and CDH and MAP datasets.

The last datasets, MAP-EM and MAP-SO, come from a list of historical forms that were not analysed or were incorrectly analysed by a previous system developed for annotating first the historical corpus CORDE[4] (Sánchez et al., 1999) and then the CDH in 2006–2009. The list was manually analysed with the aid of several specialists and contains valuable information about dating and the phenomena involved in the transformation of these old forms into modern ones. The original list contains 96 790 entries and has been split up into two lists corresponding to the Ages considered in this work.

6 Experiments and Analysis of Results

The starting point for assessing the validity of the system proposed on each of the different datasets is the performance of the lexicon derived from the DRAE using as edit transducer the identity. We will refer to this experiment establishing the baseline as ID. To have a clearer picture of the performance of the proposed system, we have carried out some experiments using the Levenshtein transducer with maximum distance costs of one and two. It is important to note that in each of the cases only the set of analysis with the lowest cost is returned. These experiments will be referred to as LEV. Finally, we have experimented with the proposed system, that will be referred to as LIN, with different maximum costs, yielding not major differences in the results. Consequently, maximum distance has been set to two. In order to compare the systems we have computed on the basis of the analysis confusion matrix the standard measures of precision, recall, their harmonic mean F, and accuracy. Formulas and results for the different datasets can be seen in Table 2.

As can be seen in Table 2, quantitatively, the LIN system obtains the best F result in all datasets, indicating the better trade-off between precision and recall, and the best accuracy rates in each dataset, while ID has better precision at CDH and LEV with a maximum distance of two obtains good results also in CDH but at the expense of overgeneration of analysis.

Most of the false negatives, i.e. missing analysis, returned by our system are due to diverging criteria regarding lemmatisation and/or categorisation which are, at least, arguable. Consider, for example, the possibility of attributing to *cerda* (sow) the masculine lemma *cerdo* (pig), as in FL-EM, or the feminine lemma *cerda* (sow), as in the other datasets. It is worth noting that false negatives caused by alternative conventions are usually accompanied by false positives,

[4] `http://corpus.rae.es/cordenet.html` (accessed 2013-03-03)

Dataset	Edit	d	TP	FP	FN	Prec.	Rec.	F	Acc.
FL-EM	ID	–	1 340	30 386	35 369	4.22	3.65	3.92	2.00
FL-EM	LEV	1	25 987	53 340	10 722	32.76	70.79	44.79	28.86
FL-EM	LEV	2	31 396	100 981	5 313	23.72	85.53	37.14	22.80
FL-EM	LIN	2	**32 679**	**14 171**	**4 030**	**69.75**	**89.02**	**78.22**	**64.23**
CDH-EM	ID	–	8 441	**2 769**	3 488	**75.30**	70.76	72.96	57.43
CDH-EM	LEV	1	10 516	5 878	1 413	64.15	88.15	74.26	59.06
CDH-EM	LEV	2	**10 805**	8 500	**1 124**	55.97	**90.58**	69.19	52.89
CDH-EM	LIN	2	10 802	3 796	1 127	74.00	90.55	**81.44**	**68.69**
CDH-SO	ID	–	25 719	**5 012**	6 249	**83.69**	80.45	82.04	69.55
CDH-SO	LEV	1	29 714	9 869	2 254	75.07	92.95	83.06	71.02
CDH-SO	LEV	2	**30 131**	12 546	**1 837**	70.60	**94.25**	80.73	**67.69**
CDH-SO	LIN	2	29 489	7 649	2 479	79.40	92.25	**85.34**	74.44
MAP-EM	ID	–	1 681	63 720	80 342	2.57	2.05	2.28	1.15
MAP-EM	LEV	1	49 825	130 974	32 198	27.56	60.75	37.92	23.39
MAP-EM	LEV	2	**63 597**	242 393	**18 426**	20.78	**77.54**	32.78	19.60
MAP-EM	LIN	2	60 201	**30 770**	21 822	**66.18**	73.40	**69.60**	**53.37**
MAP-SO	ID	–	852	31 172	39 445	2.66	2.11	2.36	1.19
MAP-SO	LEV	1	31 616	64 596	8 681	32.86	78.46	46.32	30.14
MAP-SO	LEV	2	**35 600**	92 089	**4 694**	27.88	**88.34**	42.38	26.89
MAP-SO	LIN	2	33 367	**11 779**	6 930	**73.91**	82.80	**78.10**	**64.07**

Table 2: Results of the identity transducer (ID), the Levenshtein edit transducer (LEV) with several maximum distances (d) and of the linguistic transducer (LIN) with a cost distance of two. In the table, column TP represents true positive matches, FP represents false positives, and FN false negatives. Precision $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, Recall $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, $\mathrm{F} = 2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}/(\mathrm{Precision}+\mathrm{Recall})$ and Accuracy $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$.

and that these mismatches impair both recall and precision. Some other false negatives are old word forms with no possible matches in our lexicon. They correspond to old suppletive forms or disappeared present participles. These cases are especially frequent in CDH-EM and CDH-SO and point to future work. Other false negatives found in the FL-EM dataset correspond to full Latin forms or some errata, which can not be matched by our system. For MAP datasets, there are a number of very specific formalizations that include not only morphological changes but also their combination with changes in the sound pattern and specific graphical representation found in particular texts.

As for the false positives, i.e. analysis presumably given in excess by our system, they sometimes correspond to valid lexical or morphological analysis not present in the datasets. Our lexicon contains about 4 000 old lemma variants not amenable to deactivation. Consider *faba* (bean), having both current and old uses according to DRAE. These cause false positive analysis on their word forms. It is possible to track a certain bias in both FL-EM and, much more noticeably, CDH-EM and CDH-SO datasets: CDH texts from which the lexicon is induced were manually revised and corrected, when needed, and show some particularised lemma and/or part of speech tags (or lack thereof) unattainable by our general purpose system. The lack of selective deactivation of the irrelevant fragment of the lexicon also plays an important role. Consider how *ionico* is clearly related to *jónico* (Ionian) in a medieval setting and unrelated to *iónico* (ionic), a word introduced in recent times.

Finally, regarding the core of the analysis engine, we have identified some room for improvement in the treatment of velar and sibilant orders, regressive lateral assimilation, lenition, tonal shift and rule weighting policies.

7 Conclusions and Future Work

A modular architecture for the treatment of diachronic variation has been implemented within the framework of finite-state models. The homogeneity and soundness of the framework allow different levels of linguistic variation to be easily modelled, composed and extended using previously existing resources such as modern lexicons, morphological analysers, phonological transcribers or lexical diachronic descriptions. The results obtained in all datasets show significant improvements in accuracy and in the trade-off between precision and recall with respect to the baseline and the Levenshtein distance.

Besides several adjustments to the rules, to the weights, and to the lexicon of the current system, a more in-depth qualitative analysis of the errors suggests several potential ways to further improve the performance of the system. From a lexical point of view, precision could be improved by deactivating lexicon entries corresponding to words not belonging to the ages considered. Also recall could be improved by introducing disappeared words into the lexicon and inflecting lemmas according not only to current patterns (e.g. old *andé* together with current *anduvo* (I walked)). Some changes in the morphology and morphosyntax of Spanish are not being covered by the phonological and graphical variation model (e.g. *tornarsa*, a synthetic future with a mesoclitic *tornar+se+ha* (She will come back)). This suggests the need to add a new component for this level. Finally, moving from the analysis of types to the analysis of tokens in context using language models could lead to improvements in accuracy.

Acknowledgments

We gratefully acknowledge the personnel of the *Centro de Estudios de la Real Academia Española* who worked during 2006–2009 in the CDH project from which we have created some of the datasets used in this work. We want also to thank the *Fundación Rafael Lapesa* for creating the opportunity to develop this work for the future annotation of old texts. Finally, we thank our colleagues Adelaida Fernández and Encarna Raigal for proofreading of the manuscript.

References

Allauzen, C. and Mohri, M. (2008). 3-way composition of weighted finite-state transducers. In *Proceedings of the 13th International Conference on Implementation and Application of Automata (CIAA–2008)*, pages 262–273, San Francisco, California, USA.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA–2007)*, pages 11–23, Praque, Czech Republic.

Bollmann, M., Petran, F., and Dipper, S. (2011). Applying rule-based normalization to different types of historical texts — An evaluation. In *Proceedings of the 5th Languange and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 339–344, Poznan, Poland.

Borin, L. and Forsberg, M. (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, Marrakech, Morocco.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper & Row, New York.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Jurish, B. (2010a). Efficient online k-best lookup in weighted finite-state cascades. In Hanneforth, T. and Fanselow, G., editors, *Language and Logos: Studies in Theoretical and Computational Linguistics*, volume 72 of *Studia grammatica*, pages 313–327. Akademie Verlag, Berlin.

Jurish, B. (2010b). More than words: Using token context to improve canonicalization of historical German. *Journal for Language Technology and Computational Linguistics*, 25(1):23–39.

Kaplan, R. M. and Kay, M. (1994). Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378.

Karttunen, L. (1995). The replace operator. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL–95)*, pages 16–23, Cambridge, Massachusetts, USA.

Karttunen, L. (1996). Directed replacement. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL–96)*, pages 108–115, Santa Cruz, California, USA.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Lloyd, P. M. (1987). *From Latin to Spanish*. American Philosophical Society, Philadelphia.

Mohri, M. (2009). Weighted automata algorithms. In Droste, M., Kuich, W., and Vogler, H., editors, *Handbook of Weighted Automata*, pages 213–254. Springer, Berlin.

Mohri, M. and Riley, M. (2002). An efficient algorithm for the *n*-best-strings problem. In *Proceedings of the International Conference on Spoken Language Processing 2002 (ICSLP–2002)*, Denver, Colorado, USA.

Morreale, M. (1978). Trascendencia de la *variatio* para el estudio de la grafía, fonética, morfología y sintaxis de un texto medieval, ejemplificada en el MS Esc. I.I.6. In *Annali della Facoltà di Lettere e Filosofia dell'Università di Padova*, volume II, pages 249–261, Florence, Italy.

Penny, R. J. (2002). *A history of the Spanish Language*. Cambridge University Press, Cambridge, second edition.

Piotrowski, M. (2012). Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157.

Pombo, E. L. (2012). Variation and standardization in the history of Spanish spelling. In Baddeley, S. and Voeste, A., editors, *Orthographies in Early Modern Europe*, pages 15–62. De Gruyter Mouton, Berlin, Boston.

RAE (2001). *Diccionario de la lengua española*. Espasa, Madrid, 22th edition.

Roark, B., Sproat, R., Allauzen, C., Riley, M., Sorensen, J., and Tai, T. (2012). The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the ACL 2012 System Demonstrations*, pages 61–66, Jeju Island, Korea.

Sánchez, F., Porta, J., Sancho, J. L., Nieto, A., Ballester, A., Fernández, A., Gómez, J., Gómez, L., Raigal, E., and Ruiz, R. (1999). La anotación de los corpus *CREA* y *CORDE*. In *Proceedings of SEPLN 1999*, volume 25, pages 175–182, Lleida, Spain.

Sánchez-Marco, C., Boleda, G., and Padró, L. (2011). Extending the tool, or how to annotate historical language varieties. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–9, Portland, Oregon, USA.

Wells, J. C. (1997). Sampa computer readable phonetic alphabet. In Gibbon, D., Moore, R., and Winski, R., editors, *Handbook of Standards and Resources for Spoken Language Systems*, pages 684–732. Mouton de Gruyter, Berlin and New York.

Sponsors of
NODALIDA 2013 & NEALT

The Center of Estonian
Language Resources

Mikro Værkstedet

www.tungutaekni.is

Design • Joel Priestley
Photo • Arthur Sand

NEALT Proceedings Series 18 • ISBN 978-91-7519-587-2
Linköping Electronic Conference Proceedings 87
ISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2013