A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text

Eva Pettersson[1,2], **Beáta Megyesi**[1] **and Joakim Nivre**[1]
(1) Department of Linguistics and Philology
Uppsala University
(2) Swedish National Graduate School
of Language Technology
firstname.lastname@lingfil.uu.se

Abstract

We present a multilingual evaluation of approaches for spelling normalisation of historical text based on data from five languages: English, German, Hungarian, Icelandic, and Swedish. Three different normalisation methods are evaluated: a simplistic filtering model, a Levenshtein-based approach, and a character-based statistical machine translation approach. The evaluation shows that the machine translation approach often gives the best results, but also that all approaches improve over the baseline and that no single method works best for all languages.

1 Introduction

Language technology for historical text is a field of research imposing a variety of challenges. Nevertheless, there is an increasing need for natural language processing (NLP) tools adapted to historical texts, as an aid for researchers in the humanities field. For example, the historians in the Gender and Work project are studying what men and women did for a living in the Early Modern Swedish society (Ågren et al., 2011). In this project, researchers have found that the most important words in revealing this information are verbs such as *fishing*, *selling* etc. Instead of manually going through written sources from this time period, it is therefore assumed that an NLP tool that automatically searches through a number of historical documents and presents the contained verbs (and possibly their complements), would make the process of finding relevant text passages more effective.

A major challenge in developing language technology for historical text is that historical language often is under-resourced with regard to annotated data needed for training NLP tools. This problem is further aggravated by the fact that historical texts may refer to texts from a long period of time, during which language has changed. NLP tools trained on 13th century texts may thus not perform well on texts from the 18th century. Furthermore, historical language usually shows a substantial variation in spelling and grammar between different genres, different authors and even within the same text written by the same author, due to the lack of spelling conventions.

To deal with the limited resources and the high degree of spelling variation, one commonly applied approach is to automatically normalise the original spelling to a more modern spelling, before applying the NLP tools. This way, NLP tools available for the modern language may be used to analyse historical text. Even though there may be structural differences as well between historical and modern language, spelling is the most striking difference. Moreover, language technology tools such as taggers often to some degree rely on statistics on word form n-grams and token frequencies, implying that spelling modernisation is an important step for improving the performance of such tools when applied to historical text. This paper presents an evaluation of three approaches to spelling normalisation: 1) a filtering approach based on corpus data, 2) an approach based on Levenshtein edit distance, and 3) an approach implementing character-based statistical machine translation (SMT) techniques. These approaches have previously solely been evaluated in isolation, without comparison to each other, and for one or two languages only. We compare the results of the different methods in a multilingual evaluation including five languages, and we show that all three approaches have a positive impact on normalisation accuracy as compared to the baseline. There is no single method that yields the highest normalisation accuracy for all languages, but for four out of five languages within the scope

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 32–41,
Gothenburg, Sweden, April 26 2014.

of our study, the SMT-based approach gives the best results.

2 Related Work

Spelling normalisation of historical text has previously been approached using techniques such as dictionary lookup, edit distance calculations, and machine translation.

Rayson et al. (2005) tried an approach based on dictionary lookup, where a mapping scheme from historical to modern spelling for 16th to 19th century English texts was manually created, resulting in the VARD tool (VARiant Detector) comprising 45,805 entries. The performance of the normalisation tool was evaluated on a set of 17th century texts, and compared to the performance of modern spell checkers on the same text. The results showed that between a third and a half of all tokens (depending on which test text was used) were correctly normalised by both VARD and MS Word, whereas approximately one third of the tokens were correctly normalised only when using VARD. The percentage of tokens correctly normalised only by MS Word was substantially lower; approximately 6%. VARD was later further developed into VARD2, combining the original word list with data-driven techniques in the form of phonetic matching against a modern dictionary, and letter replacement rules based on common spelling variation patterns (Baron and Rayson, 2008).

Jurish (2008) argued that due to the lack of orthographic conventions, spelling generally reflects the phonetic form of the word to a higher degree in historical text. Furthermore, it is assumed that phonetic properties are less resistant to diachronic change than orthography. Accordingly, Jurish explored the idea of comparing the similarity between phonetic forms rather than orthographic forms. For grapheme-to-phoneme conversion, a module of the IMS German Festival text-to-speech system (Black and Taylor, 1997) was used, with a rule-set adapted to historical word forms. Evaluation was performed on a corpus of historical German verse quotations extracted from *Deutsches Wörterbuch*, containing 5,491,982 tokens (318,383 types). Without normalisation, approximately 84% of the tokens were recognised by a morphological analyser. After normalisation, 92% of the tokens were recognised. Adding lemma-based heuristics, coverage increased further to 94% of the tokens.

A Levenshtein similarity approach to normalisation was presented by Bollmann et al. (2011) for Early New High German, where Levenshtein-based normalisation rules were automatically derived from 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. This normalisation method was further evaluated by Bollmann (2013), comparing the performance of the RFTagger applied to historical text before and after normalisation. For every evaluation text, the tagger was trained on between 100 and 1,000 manually normalised tokens, and evaluated on the remaining tokens in the same text. For one manuscript from the 15th century, tagging accuracy was improved from approximately 29% to 78% using this method.

Another Levenshtein-based approach to normalisation was presented by Pettersson et al. (2013b), using context-sensitive, weighted edit distance calculations combined with compound splitting. This method requires no annotated historical training data, since normalisation candidates are extracted by Levenshtein comparisons between the original historical word form and present-day dictionary entries. However, if a corpus of manually normalised historical text is available, this can optionally be included for dictionary lookup and weighted Levenshtein calculations, improving precision. This technique was evaluated for Early Modern Swedish, and in the best setting, the proportion of words in the historical text with a spelling identical to the modern gold standard spelling increased from 64.6% to 86.9%.

Pettersson et al. (2013a) treated the normalisation task as a translation problem, using character-based SMT techniques in the spelling normalisation process. With the SMT-based approach, the proportion of tokens in the historical text with a spelling identical to the modern gold standard spelling increased from 64.6% to 92.3% for Early Modern Swedish, and from 64.8% to 83.9% for 15th century Icelandic. It was also shown that normalisation had a positive effect on subsequent tagging and parsing.

Language technology for historical text also has a lot in common with adaptation of NLP tools

for handling present-day SMS messages and microblog text such as Twitter. In both genres there is a high degree of spelling variation, ad hoc abbreviations and ungrammatical structures imposing the problem of data sparseness. Similar methods for spelling normalisation may thus be used for both tasks. Han and Baldwin (2011) presented a method for normalising SMS and Twitter text based on morphophonemic similarity, combining lexical edit distance, phonemic edit distance, prefix substring, suffix substring, and the longest common subsequence. Context was taken into account by means of dependency structures generated by the Stanford Parser applied to a corpus of New York Times articles. In the best setting, a token-level F-score of 75.5% and 75.3% was reported for SMS messages and Twitter texts respectively.

3 Approaches

3.1 The Filtering Approach

The filtering approach presupposes access to a parallel training corpus of token pairs with historical word forms mapped to their modernised spelling. In the normalisation process, whenever a token is encountered that also occurred in the training data, the most frequent modern spelling associated with that token in the training corpus is chosen for normalisation. Other tokens are left unchanged.

3.2 The Levenshtein-based Approach

The Levenshtein-based approach was originally presented by Pettersson et al. (2013b). In its basic version, no historical training data is needed, which is an important aspect considering the common data sparseness issue, as discussed in Section 1. Instead, a modern language dictionary or corpus is required, from which normalisation candidates are extracted based on edit distance comparisons to the original historical word form. If there is parallel data available, i.e. the same text in its historical and its modernised spelling, this data can be used to make more reliable Levenshtein calculations by assigning weights lower than 1 to frequently occurring edits observed in the training data. The weights are then calculated by comparing the frequency of each edit occurring in the training corpus to the frequency with which the specific source characters are left unchanged, in accordance with the following formula:

$$\frac{\text{Frequency of Unchanged}}{\text{Frequency of Edit} + \text{Frequency of Unchanged}}$$

Context-sensitive weights are added to handle edits affecting more than one character. The context-sensitive weights are calculated by the same formula as the single-character weights, and include the following operations:

- double deletion: *person**ne**s* → *persons*
- double insertion: *strait* → *strai**gh**t*
- single-to-double substitution: *ju**g**e* → *ju**dg**e*
- double-to-single substitution: *m**oo**st* → *m**o**st*

For all historical word forms in the training corpus that are not identical in the modern spelling, all possible single-character edits as well as multi-character edits are counted for weighting. Hence, the historical word form *personnes*, mapped to the modern spelling *persons*, will yield weights for double-to-single deletion of *-ne*, as illustrated above, but also for single deletion of *-n* and single deletion of *-e*.

Finally, a tuning corpus is used to set a threshold for which maximum edit distance to allow between the original word form and its normalisation candidate(s). Based on the average edit distance between the historical word forms and their modern spelling in the tuning corpus, the threshold is calculated by the following formula (where 1.96 times the standard deviation is added to cover 95% of the cases):

$$\text{avg editdistance} + (1.96 * \text{standard deviation})$$

If several normalisation candidates have the same edit distance as compared to the source word, the most frequent candidate is chosen, based on modern corpus data. If none of the highest-ranked normalisation candidates are present in the corpus, or if there are several candidates with the same frequency distribution, a final candidate is randomly chosen.

3.3 The SMT-based Approach

In the SMT-based approach, originally presented by Pettersson et al. (2013a), spelling normalisation is treated as a translation task. To address changes in spelling rather than full translation of words and phrases, *character-based* translation (without lexical reordering) is performed, a well-known technique for transliteration and

character-level translation between closely related languages (Matthews, 2007; Vilar et al., 2007; Nakov and Tiedemann, 2012). In character-level SMT, phrases are modeled as character sequences instead of word sequences, and translation models are trained on character-aligned parallel corpora whereas language models are trained on character N-grams.

Since the set of possible characters in a language is far more limited than the number of possible word forms, and the same corpus will present a larger quantity of character instances than token instances, only a rather small amount of parallel data is needed for training the translation models and the language models in character-based translation. Pettersson et al. (2013a) showed that with a training and tuning set of only 1,000 pairs of historical word forms mapped to modern spelling, a normalisation accuracy of 76.5% was achieved for Icelandic, as compared to 83.9% with a full-sized training corpus of 33,888 token pairs. Their full experiment on varying the size of the training data is illustrated in Figure 1.

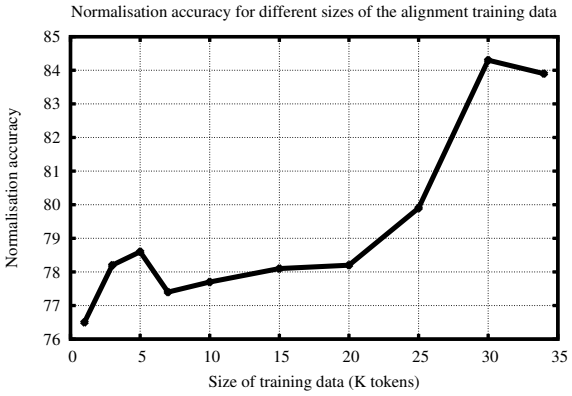

Figure 1: Normalisation accuracy when varying the size of the alignment training data.

We use the same set of training data for the SMT approach as for the filtering approach and for the assignment of weights in the Levenshtein-based approach, i.e. a set of token pairs mapping historical word forms to their manually modernised spelling. These corpora have the format of one token per line, with blank lines separating sentences. To fully adapt this format to the format needed for training the character-based translation models, the characters within each token are separated by space. The SMT system will now regard each character as a word, the full token as a sentence and the entire sentence as a section.

The SMT engine used is Moses with all its standard components. A phrase-based model is applied, where the feature weights are trained using MERT with BLEU over character-sequences as the objective function. The maximum size of a phrase (sequence of characters) is set to 10.

Two different character alignment techniques are tested: (i) the word alignment toolkit GIZA++ (Och and Ney, 2000), and (ii) a weighted finite state transducer implemented in the m2m-aligner (Jiampojamarn et al., 2007). GIZA is run with standard word alignment models for character unigrams and bigrams, whereas the m2m aligner implements transducer models based on context-independent single character and multi-character edit operations. The transducer is trained using EM on (unaligned) parallel training data, and the final model can then be used to produce a Viterbi alignment between given pairs of character strings.

An example is given in Figure 2, where the Icelandic word forms *meðr* → *meður* and *giallda* → *galda* have been aligned at a character-level using the m2m-aligner. In this example, the ϵ symbol represents empty alignments, meaning insertions or deletions. 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*. Furthermore, the alignment of *giallda* to *galda* illustrates the inclusion of multi-character edit operations, where the colon denotes a 2:1 alignment where 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: m2m character-level alignment.

4 Data

In the following, we will describe the data sets used for running the filtering approach, the Levenshtein edit distance approach, and the character-based SMT approach for historical spelling normalisation applied to five languages: English, German, Hungarian, Icelandic, and Swedish. For convenience, we use the notions of training, tun-

ing and evaluation corpora, which are well-known concepts within SMT. These data sets have been created by extracting every 9th sentence from the total corpus to the tuning corpus, and every 10th sentence to the evaluation corpus, whereas the rest of the sentences have been extracted to a training corpus.[1]

In the filtering approach, there is in fact no distinction between training and tuning corpora, since both data sets are combined in the dictionary lookup process. As for the Levenshtein edit distance approach, the training corpus is used for extracting single-character and multi-character edits by comparing the historical word forms to their modern spelling. The edits extracted from the training corpus are then weighted based on their relative frequency in the tuning corpus.

The historical texts used for training and evaluation are required to be available both in their original, historical spelling and in a manually modernised and validated spelling. A modern translation of a historical text is generally not usable, since word order and sentence structure have to remain the same to enable training and evaluation of the proposed methods. The access to such data is very limited, meaning that the data sets used in our experiments vary in size, genres and time periods between the languages.

4.1 English

For training, tuning and evaluation in the English experiments, we use the *Innsbruck Corpus of English Letters*, a manually normalised collection of letters from the period 1386–1698. This corpus is a subset of the *Innsbruck Computer Archive of Machine-Readable English Texts*, ICAMET (Markus, 1999). A subset of the British National Corpus (BNC) is used as the single modern language resource both for the Levenshtein-based and for the SMT-based approach. Table 1 presents in more detail the data sets used in the English experiments.

4.2 German

For training, tuning and evaluation in the German experiments, we use a manually normalised subset of the *GerManC* corpus of German texts from the period 1650–1800 (Scheible et al., 2011). This subset contains 22 texts from the period 1659–1780, within the genres of drama, newspaper text, letters, sermons, narrative prose, humanities, science och legal documents. The German *Parole* corpus is used as the single modern language resource both for the Levenshtein-based and for the SMT-based approach (Teubert (ed.), 2003). Table 2 presents in more detail the data sets used in the German experiments.

[1]For information on how to access the data sets used in our experiments, please contact the authors.

Resource	Data	Tokens	Types
Training	ICAMET	148,852	18,267
Tuning	ICAMET	16,461	4,391
Evaluation	ICAMET	17,791	4,573
Lev. dict.	BNC	2,088,680	69,153
Lev. freq.	BNC	2,088,680	69,153
SMT lm	BNC	2,088,680	69,153

Table 1: Language resources for English.

Resource	Data	Tokens	Types
Training	GerManC	39,887	9,055
Tuning	GerManC	5,418	2,056
Evaluation	GerManC	5,005	1,966
Lev. dict.	Parole	18,662,243	662,510
Lev. freq.	Parole	18,662,243	662,510
SMT lm	Parole	18,662,243	662,510

Table 2: Language resources for German.

4.3 Hungarian

For training, tuning and evaluation in the Hungarian experiments, we use a collection of manually normalised codices from the *Hungarian Generative Diachronic Syntax* project, HGDS (Simon, To appear), in total 11 codices from the time period 1440–1541. The Szeged Treebank is used as the single modern language resource both for the Levenshtein-based and for the SMT-based approach (Csendes et al., 2005). Table 3 presents in more detail the data sets used in the Hungarian experiments.

Resource	Data	Tokens	Types
Training	HGDS	137,669	45,529
Tuning	HGDS	17 181	8 827
Evaluation	HGDS	17,214	8,798
Lev. dict.	Szeged	1,257,089	144,248
Lev. freq.	Szeged	1,257,089	144,248
SMT lm	Szeged	1,257,089	144,248

Table 3: Language resources for Hungarian.

4.4 Icelandic

For training, tuning and evaluation in the Icelandic experiments, we use a manually normalised subset of the *Icelandic Parsed Historical Corpus* (IcePaHC), a manually tagged and parsed diachronic corpus of texts from the time period 1150–2008 (Rögnvaldsson et al., 2012). This subset contains four texts from the 15th century: three sagas (*Vilhjálm's saga*, *Jarlmann's saga*, and *Ector's saga*) and one narrative-religious text (*Miðaldaævintýri*). As a dictionary for Levenshtein calculations we use a combination of *Beygingarlýsing Íslensks Nútímamáls*, BÍN (a database of modern Icelandic inflectional forms (Bjarnadóttir, 2012)), and all tokens occurring 100 times or more in the *Tagged Icelandic Corpus of Contemporary Icelandic texts*, MÍM (Helgadóttir et al., 2012).[2] The frequency-based choice of a final normalisation candidate in the Levenshtein approach, as well as the training of a language model in the SMT approach, are done on all tokens occurring 100 times or more in the MÍM corpus. Table 4 presents in more detail the data sets used in the Icelandic experiments.

Resource	Data	Tokens	Types
Training	IcePaHC	52,440	9,748
Tuning	IcePaHC	6,443	2,270
Evaluation	IcePaHC	6,384	2,244
Lev. dict.	BÍN+MÍM	27,224,798	2,820,623
Lev. freq.	MÍM	21,339,384	9,461
SMT lm	MÍM	21,339,384	9,461

Table 4: Language resources for Icelandic.

4.5 Swedish

For training, tuning and evaluation in the Swedish experiments, we use balanced subsets of the Gender and Work corpus (GaW) of court records and church documents from the time period 1527–1812 (Ågren et al., 2011). As a dictionary for Levenshtein calculations we use SALDO, a lexical resource developed for present-day written Swedish (Borin et al., 2008). For frequency-based choice of a final normalisation candidate, we use the Stockholm Umeå corpus (SUC) of text representative of the Swedish language in the 1990s (Ejerhed and Källgren, 1997). The SUC corpus is also used to train a language model in the SMT-based approach. Table 5 presents in more detail the data sets used in the Swedish experiments.

Resource	Data	Tokens	Types
Training	GaW	28,237	7,925
Tuning	GaW	2,590	1,260
Evaluation	GaW	33,544	8,859
Lev. dict.	SALDO	1,110,731	723,138
Lev. freq.	SUC	1,166,593	97,670
SMT lm	SUC	1,166,593	97,670

Table 5: Language resources for Swedish.

5 Results

Table 6 presents the results for different languages and normalisation methods, given in terms of *normalisation accuracy*, i.e. the percentage of tokens in the normalised text with a spelling identical to the manually modernised gold standard, and *character error rate (CER)*, providing a more precise estimation of the similarity between the normalised token and the gold standard version at a character level. Table 7 summarises the results in terms of *Precision (Pre)*, *Recall (Rec)* and *F-score (F)* for the filtering approach, the Levenshtein-based approach (with and without filtering), and the best-performing SMT-based approach.

For the Levenshtein experiments, we have used context-sensitive weights, as described in Section 3.2. In the SMT approach, we run GIZA with standard word alignment models for character unigrams (un) and bigrams (bi). The m2m aligner is implemented with single character edit operations (1:1) and multi-character operations (2:2).

The baseline case shows the proportion of tokens in the original, historical text that already have a spelling identical to the modern gold standard spelling. In the Hungarian text, only 17.1% of the historial tokens have a modern spelling, with a character error rate of 0.85. For German on the other hand, accuracy is as high as 84.4%, with a character error rate of only 0.16. At a first glance, the historical spelling in the Hungarian corpus appears to be very similar to the modern spelling. A closer look however reveals recurrent differences involving single letter substitutions and/or the use of accents, as for *fiayval* → *fiaival*, *mèghalanac* → *meghalának* and *hazaba* → *házába*.

[2] The BÍN database alone is not sufficient for Levenshtein calculations, since it only contains content words.

	English		German		Hungarian		Icelandic		Swedish	
	Acc	CER	Acc	CER	Acc	CER	Acc	CER	Acc	CER
baseline	75.8	0.26	84.4	0.16	17.1	0.85	50.5	0.51	64.6	0.36
filter	91.7	0.20	94.6	0.26	75.0	0.30	81.7	0.25	86.2	0.27
Lev	82.9	0.19	87.3	0.13	31.7	0.71	67.3	0.35	79.4	0.22
Lev+filter	92.9	0.09	95.1	0.06	76.4	0.35	84.6	0.19	90.8	0.10
giza un	94.3	0.07	96.6	0.04	79.9	0.21	71.8	0.30	92.9	0.07
giza bi	92.4	0.09	95.5	0.05	80.1	0.21	71.5	0.30	92.5	0.08
m2m 1:1 un	90.6	0.11	96.0	0.04	79.4	0.21	71.2	0.31	92.3	0.08
m2m 1:1 bi	88.0	0.14	95.6	0.05	79.5	0.21	71.5	0.30	92.2	0.08
m2m 2:2 un	90.7	0.11	96.4	0.04	77.3	0.24	71.0	0.31	91.3	0.09
m2m 2:2 bi	87.5	0.14	95.5	0.05	79.1	0.22	71.4	0.31	92.1	0.08

Table 6: Normalisation results given in accuracy (Acc) and character error rate (CER).

	English			German			Hungarian			Icelandic			Swedish		
	Pre	Rec	F	Pre	Rec	F	Pre	Rec	F	Pre	Rec	F	Pre	Rec	F
filter	93.6	97.8	95.7	95.0	99.6	97.2	77.4	96.0	85.7	89.3	90.6	89.9	87.5	98.3	92.6
Lev	92.7	88.6	90.7	91.0	95.6	93.2	68.0	37.3	48.2	85.4	76.1	80.5	90.5	86.6	88.5
Lev+filter	97.4	95.2	96.3	97.3	97.7	97.5	96.2	78.8	86.7	95.6	88.0	91.7	96.6	93.8	95.2
SMT	98.2	95.9	97.0	98.7	97.9	98.3	98.3	81.3	89.0	82.0	85.2	83.6	98.6	94.1	96.3

Table 7: Normalisation results given in precision (Pre), recall (Rec) and F-score (F).

The Icelandic corpus also has a relatively low number of tokens with a spelling identical to the modern spelling. Even though the Hungarian and Icelandic texts are older than the English, German, and Swedish texts, the rather low proportion of tokens with a modern spelling in the Icelandic corpus is rather surprising, since the Icelandic language is generally seen as conservative in spelling. A closer inspection of the Icelandic corpus reveals the same kind of subtle single letter divergences and differences in the use of accents as for Hungarian, e.g. *ad* → *að* and *hun* → *hún*.

The simplistic filtering approach (filter), relying solely on previously seen tokens in the training data, captures frequently occurring word forms and works surprisingly well, improving normalisation accuracy by up to 63 percentage units. The Levenshtein-based approach (Lev) in its basic version, with no parallel training data available, also improves normalisation accuracy as compared to the baseline. However, for all languages, the simplistic filtering approach yields significantly higher normalisation accuracy than the more sophisticated Levenshtein-based approach does. This could be partly explained by the fact that frequently occurring word forms have a high chance of being captured by the filtering approach, whereas the Levenshtein-based approach runs the risk of consistently normalising high-frequent word forms incorrectly. For example, in the English Levenshtein normalisation process, the high-frequent word form *stonde* has consistently been normalised to *stone* instead of *stand*, due to the larger edit distance between *stonde* and *stand*. The even more common word form *ben*, which should optimally be normalised to *been*, has consistently been left unchanged as *ben*, since the BNC corpus, which is used for dictionary lookup in the English setup, contains the proper name *Ben*. The issue of proper names would not be a problem if a modern dictionary were used for Levenshtein comparisons instead of a corpus, or if casing was taken into account in the Levenshtein comparisons. There would however still be cases left like *stonde* being incorrectly normalised to *stone* as described above, which would be disadvantageous to the Levenshtein-based method. The low recall figures, especially for Hungarian, also indicates that there may be old word forms that are not present in modern dictionaries and thus are out of reach for the Levenshtein-based method, as for the previously discussed Hungarian word form *meghalának*.

In the Lev+filter setting, the filter is used as a first step in the normalisation process. Only tokens that could not be matched through dictionary lookup based on the training corpus are normalised by Levenshtein comparisons. The idea is

that combining these two techniques would perform better than one approach only, since high-frequent word forms are consistently normalised correctly by the filter, whereas previously unseen tokens are handled through Levenshtein comparisons. This combination does indeed perform better for all languages, and for Icelandic this is by far the most successful normalisation method of all.

For the SMT-based approach, it is interesting to note that the simple unigram models in many cases perform better than the more advanced bigram and multi-character models. We also tried adding the filter to the SMT approach, so that only tokens that could not be matched through dictionary lookup based on the training corpus, would be considered for normalisation by the SMT model. This did however not have a positive effect on normalisation accuracy, probably because the training data has already been taken care of by the SMT model, so adding the filter only led to redundant information and incorrect matches, deteriorating the results. For four out of five languages, the GIZA unigram setting yields the highest normalisation accuracy of all SMT models evaluated. For Hungarian, the GIZA bigram modell performs marginally better than the unigram model.

From the presented results, it is not obvious which normalisation approach to choose for a new language. For Icelandic, the Levenshtein-based approach combined with the filter leads to the highest normalisation accuracy. For the rest of the languages, the SMT-based approach with the GIZA unigram or bigram setting gives the best results. Generally, the Levenshtein-based method could be used for languages lacking access to annotated historical data with information on both original and modernised spelling. If, on the other hand, such data is available, the filtering approach, or the combination of filtering and Levenshtein calculations, would be likely to improve normalisation accuracy. Moreover, the effort of training a character-based SMT system for normalisation would be likely to further improve the results.

It would be interesting to also compare the results between the languages, in a language evolution perspective. This is however not feasible within the scope of this study, due to the differences in corpus size, genres and covered time periods, as discussed in Section 4.

6 Conclusion

We have performed a multilingual evaluation of three approaches to spelling modernisation of historical text: a simplistic filtering model, a Levenshtein-based approach and a character-based statistical machine translation method. The results were evaluated on historical texts from five languages: English, German, Hungarian, Icelandic and Swedish. We see that all approaches are successful in increasing the proportion of tokens in the historical text with a spelling identical to the modernised gold standard spelling. We conclude that the proposed methods have the potential of enabling us to use modern NLP tools for analysing historical texts. Which approach to choose is not clear, since the results vary for the different languages in our study, even though the SMT-based approach generally works best. If no historical training data is available, the Levenshtein-based approach could still be used, since only a modern dictionary is required for edit distance comparisons. If there is a corpus of token pairs with historical and modern spelling available, training an SMT model could however result in improved normalisation accuracy. Since the SMT models are character-based, only a rather small amount of training data is needed for this task, as discussed in Section 3.3.

We believe that our results would be of interest to several research fields. From a language evolution perspective, future research would include a thorough investigation of why certain approaches work better for some languages but not for other languages, and what the results would be if the data sets for the different languages were more similar with regard to time period, size, genre etc. The latter could however be problematic, due to data sparseness. For historians interested in using modern NLP tools for analysing historical text, an extrinsic evaluation is called for, comparing the results of tagging and parsing using modern tools, before and after spelling normalisation. Finally, the proposed methods all treat words in isolation in the normalisation process. From a language technology perspective, it would be interesting to also explore ways of handling grammatical and structural differences between historical and modern language as part of the normalisation process. This would be particularly interesting when evaluating subsequent tagging and parsing performance.

References

Maria Ågren, Rosemarie Fiebranz, Erik Lindberg, and Jonas Lindström. 2011. Making verbs count. The research project 'Gender and Work' and its methodology. *Scandinavian Economic History Review*, 59(3):271–291. Forthcoming.

Alistair Baron and Paul Rayson. 2008. Vard2: A tool for dealing with spelling variation in historical corpora. In *Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham.

Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. In *AfLaT2012/SALTMIL joint workshop on Language technology for normalisation of less-resourced languages*, Istanbul, May.

Alan W. Black and Paul Taylor. 1997. Festival speech synthesis system: system documentation. Technical report, University of Edinburgh, Centre for Speech Technology Research.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 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.

Marcel Bollmann. 2013. POS tagging for historical texts with sparse training data. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 11–18, Sofia, Bulgaria, August. Association for Computational Linguistics.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. Saldo 1.0 (svenskt associationslexikon version 2). Språkbanken, University of Gothenburg.

C. Csendes, J. Csirik, T. Gyimóthy, and A. Kocsor. 2005. The Szeged Treebank. In *Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005)*, Karlovy Vary, Czech Republic.

Eva Ejerhed and Gunnel Källgren. 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.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In Association for Computational Linguistics, editor, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 368–378, Portland, Oregon, USA, June.

Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir, and Hrafn Loftsson. 2012. The Tagged Icelandic Corpus (MÍM). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages*, pages 67–72.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 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, April.

Bryan Jurish. 2008. Finding canonical forms for historical German text. In Angelika Storrer, Alexander Geyken, Alexander Siebert, and Kay-Michael Würzner, editors, *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing* (KONVENS 2008), pages 27–37. Mouton de Gruyter, Berlin.

Manfred Markus, 1999. *Manual of ICAMET (Innsbruck Computer Archive of Machine-Readable English Texts)*. Leopold-Franzens-Universität Innsbruck.

David Matthews. 2007. Machine transliteration of proper names. Master's thesis, School of Informatics.

Preslav Nakov and Jörg Tiedemann. 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, July. Association for Computational Linguistics.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. pages 440–447, Hongkong, China, October.

Eva Pettersson, Beáta Megyesi, and Tiedemann Jörg. 2013a. An SMT approach to automatic annotation of historical text. In *Proceedings of the NoDaLiDa 2013 workshop on Computational Historical Linguistics*, May.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2013b. Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference on Computational Linguistics (NoDaLiDa)*, May.

Paul Rayson, Dawn Archer, and Nicholas Smith. 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, July.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurdsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. A Gold Standard Corpus of Early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 124–128, Portland, Oregon, USA, June. Association for Computational Linguistics.

Eszter Simon. To appear. Corpus building from Old Hungarian codices. In Katalin É. Kiss, editor, *The Evolution of Functional Left Peripheries in Hungarian Syntax*. Oxford University Press.

Wolfgang Teubert (ed.). 2003. German Parole Corpus. Electronic resource, Oxford Text Archive.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague, Czech Republic, June. Association for Computational Linguistics.