Using Comparable Collections of Historical Texts for Building a Diachronic Dictionary for Spelling Normalization

Marilisa Amoia
Saarland University
m.amoial@mx.uni-saarland.de

Jose Manuel Martinez
Saarland University
j.martinez@mx.uni-saarland.de

Abstract

In this paper, we argue that comparable collections of historical written resources can help overcoming typical challenges posed by heritage texts enhancing spelling normalization, POS-tagging and subsequent diachronic linguistic analyses. Thus, we present a comparable corpus of historical German recipes and show how such a comparable text collection together with the application of innovative MT inspired strategies allow us (i) to address the word form normalization problem and (ii) to automatically generate a diachronic dictionary of spelling variants. Such a diachronic dictionary can be used both for spelling normalization and for extracting new "translation" (word formation/change) rules for diachronic spelling variants. Moreover, our approach can be applied virtually to any diachronic collection of texts regardless of the time span they represent. A first evaluation shows that our approach compares well with state-of-art approaches.

1 Introduction

The study of heritage documents has been one of the regular sources of knowledge in the Humanities, specially in history-related disciplines. The last years have witnessed an increased interest in approaches combining NLP and corpus-based techniques in the Humanities (Piotrowski, 2012) as they can provide new insights and/or a more consistent and reliable account of findings.

Until recently, research efforts have been focused on building diachronic corpora (e.g. Old Bailey Online project (Hitchcock et al., 2012) and its follow-up, the Old Bailey Corpus (Huber, 2007), the Bonn Corpus of Early New High German (Diel et al., 2002) or the GerManC (Scheible et al., 2011b) for German and many others). Such resources are generally annotated with shallow metadata (e.g. year of publication, author, geographical location) for allowing fast retrieval. However, the annotation of richer linguistic and semantic information still poses a series of challenges that have to be overcome, such as (i) the noise introduced by deviant linguistic data (spelling/orthography variation, lack of sentence boundaries, etc.) typical of this kind of material, due to the lack of standardized writing conventions in terms of words and punctuation and hence (ii) the higher error rates obtained when applying standard NLP methods.

Further, standardization of spelling variation in historical texts can be broken down at least into two subproblems:

1. the old word forms often differ from the modern orthography of the same items. Consider, for instance, the diachronic variants of the third person singular of present tense of the verb *werden* in German (which means 'become' as full verb, or is used as auxiliary verb to build the future): wirt, wirdt, wirdet vs wird; (Piotrowski, 2012) and

2. the denomination of certain objects may result completely different from that used in the modern language due to historical reasons (e.g. adoption of foreign language terms, semantic shift). Consider, as an example, the German historical/modern variants of the word *lemon* (e.g. *Limonie*/Zitrone) or of the word *woman* (e.g. *Weib*/Frau).

Previous approaches to spelling normalization of historical texts have focused on the first subproblem. Two main strategies that have been applied:

- a rule based strategy, in which the translation of historical variants into modern forms

Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 84–89,
Sofia, Bulgaria, August 8 2013. ©2013 Association for Computational Linguistics

is performed on the ground of manually written or semi-automatically gathered rules (cf. (Pilz et al., 2008), (Bollmann et al., 2011));

- a string similarity strategy, in which a semi-automatic attempt is made to link historical variants with modern dictionary entries following string similarity (cf. (Giusti et al., 2007), (Kunstmann and Stein, 2007), (Dipper, 2010), (Hendrickx and Marquilhas, 2011), (Gotscharek et al., 2011)) or phonetic conflation strategies (cf. (Koolen et al., 2006), (Jurish, 2008)).

These approaches have the disadvantage of ending up relying on a time-specific dictionary of variants, e.g. they can cope with variants realized in texts stemming from the same period of time for which they have been created but may result inappropriate for texts belonging to other time spans.

Moreover, to our knowledge, there is currently no approach to spelling normalization that can address successfully the second subproblem stated above – the recognition of paraphrastic variations realized as completely different strings or consisting of semantic shifts.

As it has been often noted, the problem of standardizing diachronic variants can be understood as a translation operation, where instead of translating between two different languages, translation takes place between two diachronic varieties of the same language. Inspired by experiments done for interlinguistic translation (Rapp et al., 2012), the idea is to use diachronic comparable corpora to automatically produce a dictionary of diachronic spelling variants even including semantic shifts, regardless of the historical variants at stake.

In short, we first build a comparable historical corpus made up of recipe repertoires published in the German language during the Early Modern Age along with a contemporary comparable corpus. Second, we address the problem of recognizing and translating different variants by relaying on MT techniques based on string similarity as well as on semantic similarity measures. Finally, we automatically extract a diachronic dictionary of spelling and semantic variants which also provides a canonical form.

This paper is organized as follows. Section 2 presents the comparable corpus of German recipes. Section 3 describes the approach used for generating the dictionary of diachronic spelling variants. Section 4 shows the results of a preliminary evaluation. Finally, in Section 5 we conclude by discussing some final remarks.

2 The Historical Comparable Corpus of German Recipes

The text collection encoded in our corpus spans two hundred years and includes samples from 14 cook books written in German between 1569 and 1729. The core of the recipe corpus was compiled as part of a PhD work in the field of Translation Studies (cf. (Wurm, 2007)). This corpus has been aligned resulting into two comparable corpora:

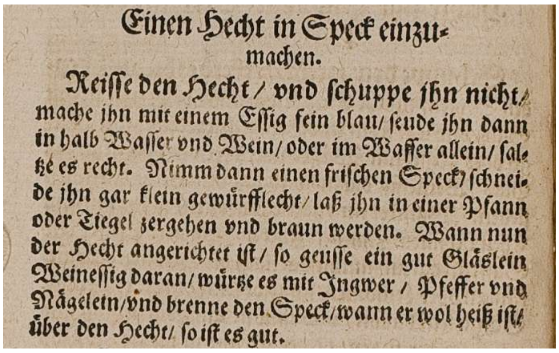
- a historical comparable dataset aligned at recipe level providing multiple versions of the same dish across the time span of the core corpus;
- a contemporary comparable dataset providing contemporary German versions for each recipe.

In order to produce the historical comparable component we proceeded in the following way:

- first, we classified the core recipes by main ingredient and cooking method (e.g. chicken, roast). These two parameters form the criteria to consider the recipes aligned, then;
- we collected as many as possible diachronic versions/variants of the same recipe by also searching online resources providing collections of historical texts.

The historical component of the corpus (core and comparable) contains a total of 430 recipes and about 45.000 tokens. This dataset constitutes the object of study for subsequent research, providing a representative sample of German during the Early Modern Age in this specific domain. Moreover, language and genre evolution can be traced thanks to its comparable nature.

Regarding the compilation of the contemporary German comparable corpus, we collected a set of recipes belonging to the same register but representing contemporary German language. These recipes were collected from Internet sources and filtered by geographical criteria (only the ones categorized as belonging to the cuisine of German speaking regions were selected). The corpus contains around 1500 recipes and over 500.000 tokens, which have been also aligned with their



Einen Hecht in Speck einzu-
machen.
Reiſſe den Hecht / vnd ſchuppe jhn nicht/
mache jhn mit einem Eſſig fein blau/ ſeude jhn dann
in halb Waſſer vnd Wein/ oder im Waſſer allein/ ſal-
tze es recht. Nimm dann einen friſchen Speck/ ſchnei-
de jhn gar klein gewürfflecht/ laß jhn in einer Pfann
oder Tiegel zergehen vnd braun werden. Wann nun
der Hecht angerichtet iſt/ ſo geuſſe ein gut Gläslein
Weineſſig daran/ würtze es mit Jngwer / Pfeffer vnd
Nägelein/ vnd brenne den Speck/ wann er wol heiß iſt/
über den Hecht/ ſo iſt es gut.

Figure 1: A text excerpt from Wecker 1679.

comparable historical counterparts according to the same parameters explained above. This subset allows not only to compare historical recipes with their modern versions but also to use them as a reference corpus to extract standard word forms.

2.1 Digitization Strategy

The corpus has been manually transcribed. The transcription can be regarded as a diplomatic one, since it tries to preserve as many features of the original as possible. Some standardization has been performed at punctuation and hyphenation level but no spellchecking or word separation has been carried out. The corpus is encoded in UTF-8 and we have used a TEI-compatible XML format to store both text and metadata.

2.2 Annotations

The corpus currently includes some shallow semantic annotation describing text structure (e.g. recipe, title, and body) and providing a basic classification of recipes based on the main ingredient and recipe type. The figure 2 below shows an example of semantic annotation.

```
<recipe id=”26” author=”Deckhardt” year=”1611”
language=”german” ingredient=”Erdbeere”
cookingMethod=”Mus”>
<title> Ein Erdbeermuhs zumachen. < /title>
<body> <seg type=”newline”>
Nimb Erdbeer
</seg>
<seg type=”newline”/ >
treibe es durch mit Weine
</seg>
<seg type=”newline”>
thue Zucker darein
</seg>
<seg type=”newline”>
darnach man es gerne süsse haben wil
</seg>
...
< /body>
< /recipe>
```

Figure 2: Comparable diachronic corpus: an example of annotation.

3 Building a Diachronic Dictionary of Spelling Variants

Our spelling normalization strategy aims at solving both subproblems discussed in the Introduction. In order to extract the mapping between diachronic variants by also capturing paraphrases and semantic shifts, we apply two different strategies one based on string similarity and the other based on semantic similarity measures.

Our workflow can be summarized as follows:

1. In a first step, we relay on clustering techniques based on string similarity measures to identify a set of diachronic variations of the same word form. In this phase, clustering corresponds to the extraction of ”similar strings”.

2. In the second step, we address the problem of finding semantic variants, i.e. those variants that are not realized as similar strings by applying paraphrase recognition techniques to identify different denominations of the same object.

3. Finally, we integrate the results of both phases and generate a dictionary of diachronic variants, that is used to extract the normalized spelling for each word in the corpus. We assume that the normalized word form corresponds to the most modern variant found in the dictionary.

3.1 String Similarity

In the first step, we extract comparable recipes from different decades and from the corpus of modern recipes. Then we apply clustering techniques to find spelling variations. The fact that we use comparable texts for clustering, should reduces the errors as all tokens come from similar terminological fields.

We apply agglomerative hierarchical clustering as implemented in the R statistical programming environment with the *average* agglomeration method. As a string similarity measure, we use the standard Levenshtein edit distance as implemented in the R package *Biostrings*. In order to build the dictionary, we select clusters that have a string similarity greater than 65%. Figure 3 shows an example of diachronic dictionary entries generated with this approach.

Hühner:	Hüner_1574, Hünern_1574, hüner_1574, Hünner_1611
und:	vnd_1569, vnnd_1569, vnd_1679, und_1698
magsts:	magst_1574, magstu_1602, magst_1679
lasst:	lassen_1679, lassets_1682, lässets_1715
Muscatenblüh:	Muscatblü_1569, Muscatenblüh_1715

Figure 3: Diachronic Dictionary.

For each list of diachronic variants gathered at this point, we extracted the most recent variant and used it as normalized form.

3.2 Semantic similarity

In order to cluster paraphrastic variants and semantic shifts, we apply a slightly modified version of Lin's algorithm (Lin, 1998) based on the assumption that words sharing similar contexts should have similar semantics. Contrary to Lin, in our approach we do not perform any dependency analysis of the corpus data and compute semantic similarity between strings simply in terms of the mutual information of trigrams.

The semantic similarity strategy we implemented can be summarized as follows:

- We start by generating a list of trigrams from the corpus.
- We assign to each pair of tokens in the corpus a value for their mutual information.
- We assign to each pair of tokens in the corpus a value for their similarity.
- For each token in the corpus, we extract the N most similar tokens and take the most modern one as the normalized form.

The mutual information I for a pair of tokens $t1$ and $t2$ is defined as:

$$I(t1,t2) = log\frac{\|t1,*,t2\|\|*,*,*\|}{\|t1,*,*\|\|*,*,t2\|}, \text{ with}$$

$\| t1,*,t2 \|$ the frequency of the occurrence of the trigram *t1,*,t2* in the corpus, $\| *,*,* \|$ the total number of trigrams in the corpus, $\| t1,*,* \|$ the number of trigrams with $t1$ as first token and $\| *,*,t2 \|$ the number of trigrams with $t2$ as last token.

Semantic similarity between tokens is defined in terms of their mutual information:

$$sim(t1,t2) = \frac{\sum_{T_{t1} \cap T_{t2}} I(t1,*)+I(t2,*)}{\sum I(t1,*)+\sum I(w2,*)},$$

with $T_{t1} = \{(v,w) : I(t1.w) > 0\}$ and $T_{t2} = \{(v,w) : I(t2.w) > 0\}$, the sets of token pairs that form trigrams with t1 or t2 as first element and such that they have positive mutual information values.

4 Evaluation

In order to evaluate the performance of our normalization strategy, we extracted a subset of recipes from the corpus for testing purposes. This subcorpus includes 32 comparable recipes on how to roast a chicken that have been written in a time period ranging from 1569 to 1800 reaching a size of 7103 words (about 8% of whole corpus). We take as reference the results yielded by TreeTagger[1] (Schmid, 1994), the state-of-art POS-tagger for German, regarding lemmatization and POS-tagging.

First, we tagged the subcorpus on the non-normalized word forms. The performance of POS-tagging, in this case, is around 80%, which is higher than the one observed in similar experiments (cf. (Scheible et al., 2011a)) on other historical corpora of German. We believe the reason for this is the relative syntactic simplicity of recipe texts in comparison to other kind of texts (dramas, sermons, letters, scientific or legal texts).

The tagger's poor performance is due to the existence of lexical items unknown to the system (around 27%), on the one hand, and the high inconsistency of the spelling, on the other hand. Our strategy to circumvent this problem consists of providing a modern word form to all historical word variants that we obtained from the previously discussed diachronic dictionary. We expected, that after the two normalization steps discussed in Section 3, the performance of the tagging process should improve.

[1] The TreeTagger was trained on the TüBa-D/Z treebank. Its performance is about 97.4% on newspaper texts and 78% on texts containing unknown words.

Strategy	Lemma	POS
no-norm	73%	80%
string-similarity	81%	81.4%
semantic similarity	82.5%	82%

Table 1: Evaluation Results.

Therefore, we repeated the experiment, first, on the test subcorpus normalized by using the diachronic dictionary generated with first normalization strategy, i.e. the one based on string similarity measure and, second, on the normalized version obtained after using the second strategy based on semantic similarity.

Table 1 summarizes the results of a preliminary evaluation of our strategy.

After string similarity normalization, the tagger was able to identify all lemmas except for 1358 tokens (19% of unknown tokens). While POS-tagging improved to 81.4%.

The semantic similarity step improved the performance of lemmatization and POS reaching 82.5% and 82% respectively.

Despite the fact that our experiments refer to very few data and to a restricted domain, we believe they are promising and show that our strategy, the integration of string similarity and semantic similarity measures can lead to a high quality automatic spelling normalization and outperform state-of-art approaches.

5 Conclusion

In this paper we have presented a comparable corpus of historical German recipes and shown that such comparable resources can help removing sources of noise typical of these text types that hinder standard NLP manipulation of such material. The old German recipes corpus is, to our knowledge, one of the first attempts[2] to build a comparable historical corpus of German. The corpus is accessible through a web interface and allows sophisticated queries according to different levels of annotation: 1) historical word forms; 2) modern normalized forms; 3) lemmas on top of normalized forms; 4) part-of-speech, and, last but not least; 5) semantics, namely main ingredient and cooking method. Further, we describe an innovative strategy for word form normalization that integrate string similarity measure with semantic similarity thereby being able to cope not only with formal spelling variations but also with paraphrastic variations and semantic shift. Moreover, this method can be applied to any comparable diachronic corpus, regardless of the time span at stake. A preliminary evaluation has shown that such a strategy may outperform state-of-art approaches.

[2] We are aware of only one similar project (Bartsch et al., 2011) aimed at building a comparable corpus of German texts for three main periods Old High, Middle High and Early New High German. However, those corpora are not yet available.

References

Nina Bartsch, Stefanie Dipper, Birgit Herbers, Sarah Kwekkeboom, Klaus-Peter Wegera, Lars Eschke, Thomas Klein, and Elke Weber. 2011. Annotiertes Referenzkorpus Mittelhochdeutsch (1050-1350). Poster session at the 33rd annual meeting of the German Linguistic Society (DGfS-2011) (Abstract, Poster) .

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Applying Rule-Based Normalization to Different Types of Historical Texts — An Evaluation. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, November.

Marcel Diel, Bernhard Fisseni, Winfried Lenders, and Hans-Christian Schmitz. 2002. XML-Kodierung des Bonner Frühneuhochdeutschkorpus. IKP-Arbeitsbericht NF 02, Bonn .

Stefanie Dipper. 2010. Pos-tagging of historical language data: First experiments. In *Semantic Approaches in Natural Language Processing. Proceedings of the 10th Conference on Natural Language Processing (KONVENS-10)*, pages 117–121, Saarbrücken.

Rafael Giusti, Arnaldo Candido Jr, Marcelo Muniz, Lívia Cucatto, and Sandra Aluísio. 2007. Automatic Detection of Spelling Variation in Historical Corpus : An Application to Build a Brazilian Portuguese Spelling Variants Dictionary. In *Proceedings of the Corpus Linguistics Conference*, pages 1–20.

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

Iris Hendrickx and Rita Marquilhas. 2011. From old texts to modern spellings: an experiment in automatic normalisation. *Journal for Language Technology and Computational Linguistics*, 26(2):65–76.

Tim Hitchcock, Robert Shoemaker, Clive Emsley, Sharon Howard, and Jamie McLaughliin. 2012. The Old Bailey Proceedings Online, 1674-1913 (version 7.0).

Magnus Huber. 2007. The Old Bailey Proceedings, 1674-1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English. In

Meurman-Solin. Anneli and Arja Nurmi, editors, *Annotating Variation and Change*, volume 1. Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki, Helsinki.

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 27–38. Mouton de Gruyter, Berlin / New York.

Marijn Koolen, Frans Adriaans, Jaap Kamps, and Maarten de Rijke. 2006. A cross-language approach to historic document retrieval. In Mounia Lalmas, Andy MacFarlane, Stefan Rueger, Anastasios Tombros, Theodora Tsikrika, Alexei Yavlinsky, editor, *Advances in Information Retrieval*, volume 3936, pages 407–419. Lecture Notes in Computer Science, Berlin/Heidelberg: Springer.

Pierre Kunstmann and Achum Stein. 2007. Le Nouveau Corpus d'Amsterdam. In Pierre Kunstmann Achim Stein, editor, *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*, pages 9–27. Stuttgart, Germany: Steiner.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words.

Thomas Pilz, Andrea Ernst-Gerlach, Sebastian Kempken, Paul Rayson, and Dawn Archer. 2008. The Identification of Spelling Variants in English and German Historical Texts: Manual or Automatic? *Literary and Linguistic Computing*, 23(1):65–72, April.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*, volume 5. Morgan & Claypool Publishers, September.

Reinhard Rapp, Serge Sharoff, and Bogdan Babych. 2012. Identifying word translations from comparable documents without a seed lexicon. In *LREC*, pages 460–466.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011a. Evaluating an f-the-shelfS-tagger on Early Modern German text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, number June, pages 19–23, Portland, Oregon. Association for Computational Linguistics.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011b. A gold standard corpus of early modern german. In *Linguistic Annotation Workshop*, pages 124–128.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Andrea Wurm. 2007. *Translatorische Wirkung: ein Beitrag zum Verständnis von Übersetzungsgeschichte als Kulturgeschichte am Beispiel deutscher Übersetzungen französischer Kochbücher in der Frühen Neuzeit.* Ph.D. thesis, Universität des Saarlandes, Saarbrücken.