Barteld, Fabian, Ingrid Schröder und Heike Zinsmeister. 2015. Unsupervised regularization of historical texts for POS tagging. In: Proceedings of the 4th Workshop on Corpus-based Research in the Humanities (CRH), Warschau, Polen, 3-12.

Unsupervised regularization of historical texts for POS tagging

Fabian Barteld, Ingrid Schröder and Heike Zinsmeister

Institut für Germanistik Universität Hamburg E-mail: firstname.lastname@uni-hamburg.de

Abstract

This paper presents an unsupervised method to reduce spelling variation in historical texts in order to mitigate the problem of data sparsity. In contrast to common normalization techniques, the historical types are not mapped to corresponding types in a standardized target language (e.g. normalizing Early New High German to Modern German). Consequently, no additional resources such as manually normalized data, parallel texts or a dictionary of the target language are needed. Furthermore, our approach does not use any annotation and is thus not dependent on the existence of annotated data. We evaluate the usefulness of this approach using POS tagging.

1 Introduction

In the DFG-funded project *Reference Corpus Middle Low German/Low Rhenish* (1200-1650), a corpus of Middle Low German (GML) and Low Rhenish texts annotated with fine-grained parts of speech and lemmas is created for linguistic and philological research.¹ The annotation must be of high quality for the corpus to be accepted as a reference corpus by relevant research communities. In order to speed up the annotation process and to ensure consistent annotation, automatic preannotations and error detection methods are used. However, the initial usefulness of statistical approaches is limited due to data sparsity. GML is a low-resourced historical dialect and it exhibits a great deal of spelling variation, as is common for historical texts; furthermore, the corpus contains texts from different time periods, dialect regions and domains, which amplifies the spelling variation in the overall corpus.

This paper contributes to a line of research that uses normalization methods to overcome the problem of data sparsity induced by spelling variations when training

¹The project is a cooperation between the University of Hamburg (Ingrid Schröder) and the University of Münster (Robert Peters). More information about the project can be found at http://referenzkorpus-mnd-nrh.de.

a POS tagger on historical texts. We present an unsupervised technique to reduce spelling variation that requires no target language – unlike common normalization techniques used for historical texts (e.g. normalizing Early New High German to Modern German) – nor does it use any annotation. Consequently, this technique is not dependent on additional language resources such as manually normalized or annotated data, the existence of parallel texts or a dictionary of the target language.

2 Related work

Spelling variation is problematic for statistical NLP applications, as it increases data sparsity. This issue can either be addressed by adapting the NLP tools to handle the variation or by a pre-processing step that reduces the variation before an off-the-shelf implementation of a tool is applied to the data. Here, we concentrate on the second approach, often called *normalization*.

Most of the work on normalizing historical texts uses a standardized modern language as a target and treats normalization as some kind of "transformation of a historical form into its modern equivalent" [2]. One problem with normalization techniques of this kind – including unsupervised ones – is that they always require additional resources to define the target language. These may be parallel texts used to create training pairs of historical variant and target language form [3], an unannotated corpus and a lexicon of the target language [17] or, at the least, a lexicon that defines the target language [13]. This is unproblematic for historical variants that are closely related to a standardized modern language. In the case of GML, the problem is that there is no such standardized language. The modern version of GML – Low German – is still merely a group of spoken dialects without a standardized written form. For this reason, we are pursuing a different approach that we call *regularization*² to distinguish it from normalization with a target language.³

Regularization can be defined as the *conflation* of all spelling variants into one target form. Normalization as defined above is then a special instantiation of regularization in which the target forms are defined externally by a target language.⁴

In [14] and [16], two approaches are presented whereby possible spelling variants of one word are induced directly from a historical corpus in an unsupervised fashion. However, both of these approaches rely on existing annotations (lemmas [14] or POS tags [16]). As we are creating the first annotated corpus for GML, the amount of annotated data available is very limited. This motivates us to explore an approach toward regularization that uses tokenized, unlabelled texts as input without any additional resources for the selection of conflation candidates in the texts.

²We would like to thank the anonymous reviewers for their comments on the terminology.

³Note that there are similar approaches that do not make this terminological distinction, e.g. [16] refer to normalization without using a target language.

⁴However, in cases in which the target language exhibits spelling variation as well (e.g. the variation between *-s* and *-es* as the genitive marker in Modern German), normalization differs from regularization as spelling variants might be mapped to different target forms in the former approach.

Normalization is used not only to reduce the spelling variation in a corpus but also to enable the general usage of existing resources for the target language. This re-usage of existing resources is not possible when using regularization. However, there are many corpus-related tasks that do not depend on additional resources that can benefit from the reduction of spelling variation, e.g. keyword statistics [1] and automatic error-detection with systems like the one described in [6].

We evaluate our approach by testing the impact of regularization on the accuracy of POS tagging in GML texts. Related work has shown that the POS tagging of Middle High and Early New High German texts can be improved by reducing the spelling variation. In [7], Dipper compares the tagging of diplomatic transcriptions of Middle High German texts with that of normalized variants for which the normalization was realized with handwritten rules and manually corrected. She reports improvements in per-word accuracy for POS tagging between 3.75% and 4.81% for different parts of the corpus. In [16], Logačev et al. conflate possible spelling variants before training and applying a POS tagger on Early New High German texts. They report mixed results with a maximum of 1.6% improvement in per-word accuracy, including a decrease in accuracy for one text (-0.2%).

3 The data

In this study, we use four texts (see Table 1; full bibliographical information is given in the bibliography) from the GML Reference Corpus that have been manually corrected for tokenization and sentence boundaries. We employ a simplified version of the transcription in which abbreviations are expanded, among other aspects, thereby already reducing spelling variation. For the experiments, all tokens have been lowercased.

Name	Year	Domain	Туре	Tokens	Types
Johannes	~1480	religious texts	manuscript	19645	2305
Griseldis	1502	literature	print	9062	2251
OldenbSSP	1336	law	manuscript	21800	2731
SaechsWeltchr	1 st half 14 th c.	arts	manuscript	18215	3255

Table 1: The texts used for the experiments

All texts are from the same dialect region (North Low Saxon). They only differ in the year of writing/printing, the textual domain and the medium (print or manuscript). As alluded to above, they exhibit a large number of spelling differences, e.g. the 3SG.PST.IND of the verb *blîven* '(to) stay' appears as *blef* (6x) in Johannes and as *bleff* (3x) in Griseldis (see Examples 1 and 2).

(1) vnde nemande bleff vngewenet . and nobody.SG.NOM stay.3SG.PST.IND NEG-PTCP-cry-PTCP . 'and nobody could help but crying' (Griseldis)

(2) vnde blef vp eme and stay.1SG.PST.IND upon he.3SG.M.DAT 'and stayed by him' (Johannes)

Such spelling differences are not limited to instances between texts. Spelling variation is also observed within individual texts, as can be seen with the example *anbegin* 'beginning', which appears as *anbegin* and as *anbegyn* in Griseldis and in a third version (*anbeginne*) in Johannes. In the current experiment, we ignore spelling variation that appears within one text and concentrate on regularization of the variation between texts.

For the evaluation of the POS tagging, Johannes and Griseldis (cf. Table 1) have been manually tagged with the POS tagset HiNTS, which is an adapted version of HiTS [8], a tagset created for the GML reference corpus that consists of 105 tags.⁵ Both texts contain a comparable number of types, although Johannes is more than twice as long as Griseldis. The reason for this is that the gospel of Johannes is a rather repetitive and formulaic text. This is reflected in the better POS-tagging accuracy results for Johannes in a 20-fold cross-evaluation (with whole sentences) on each text individually using RFTagger⁶ [19]. The accuracy is 90.0% \pm 1.5 for Johannes and 83.9% \pm 2.7 for Griseldis. When training the tagger only with "outof-domain" data, (i.e. on the other text), we observe for both texts a drop in the mean accuracy of about 15% (cf. Table 2) on the same 20 parts.

	% correct	% correct (known)	% correct (unknown)	% unknown
Johannes	73.5 ± 1.9	83.2 ± 2.4	48.7 ± 3.5	28.3 ± 3.1
Griseldis	69.3 ± 3.3	80.6 ± 2.3	46.6 ± 6.6	33.0 ± 3.3

Table 2: Per-word tagging accuracy trained on out-of-domain data

The differences in performance can be easily explained by the amount of unknown words, which is higher when only "out-of-domain" data is used for training and which is also higher for Griseldis in general. As some of the unknown words are due to spelling variations, we expect that substituting an unknown type with a known spelling variant will improve the performance of the POS tagger.

⁵At the time of writing, the GML corpus is still under construction, and the POS annotation used for evaluation in our experiments is pre-final.

⁶RFTagger performed second-best in initial tests, slightly below HunPos [10]. Both taggers outperformed SVMTool [9] and CRFSuite [18] with standard POS-tagging features. RFTagger has been chosen over HunPos for the presented experiments, as morphological categories are added to the POS tags in further steps. However, our regularization technique improved POS-tagging accuracy for all of these taggers.

4 Regularization with string and context similarity

In this section, we present a language-independent approach to unsupervised regularization using string and context similarity.

For string similarity, we use the similarity measure *Proxinette* [11, 12]. Proxinette was designed to measure the morphological similarity of lexemes and to find morphologically related words in a lexicon. We use Proxinette in the variant described in [12], where character n-grams are the only features.

The intuition behind Proxinette is that the more character n-grams two types share, the more similar they are. The n-grams are weighted by their frequency in the corpus: More frequent n-grams contribute less to the similarity. In the original version of Proxinette, the character n-grams have a minimum length of three. However, this leaves spelling variants such as *yck* and *ic* 'I' unconnected. Therefore, we vary the minimal n-gram length as a parameter in our experiments (Ngram). Proxinette is computed based on a graph and is consequently efficient and scalable and can be employed to compare many types. Proxinette returns the similarity of two types as a number between 0 and 1: 0 means no similarity (i.e. no shared n-gram), while higher values denote a greater similarity.

We use Proxinette similarity to select the known types that are most similar (i.e. have the maximal Proxinette similarity of all known types) as conflation candidates for an unknown type. We restrict this choice by a threshold for the similarity, which is also varied as a parameter in our experiments (Prox). A higher threshold reduces the overall number of types for which conflation candidates are generated.

The conflation candidates are then filtered using Brown clusters [4], allowing only the candidates chosen by Proxinette that fall into the same cluster. The number of Brown clusters is varied as an additional parameter [5] (Brown). When more than one conflation candidate exists, one is chosen at random.

For Proxinette, we use our own implementation. The Brown clusters are computed using the implementation described in [15].⁷ As a fourth parameter, we vary the amount of data utilized to compute the Proxinette similarity and the Brown clusters (Data): We either use only Johannes and Griseldis (base) or all texts presented in Table 1 (all).

5 Results

In this section, we present the results of a POS-tagging experiment using the regularization technique described in the previous section. The baseline is given by tagging the raw texts (cf. Table 2). As with the baseline, we tag the texts Johannes and Griseldis with RFTagger trained on the other text. In contrast to the baseline, the text to be tagged is first regularized by conflating unknown types with similar known types.

⁷The source code is available at https://github.com/percyliang/brown-cluster.

	Text	Ngram	Prox	Brown	Data	% correct
best	Johannes	1	0.000001	125	all	75.7 ± 1.6
	Griseldis	1	0.02	25	base	71.4 ± 3.0
combined	Johannes	2	0.000001	50	all	75.6 ± 1.7
best	Griseldis	2	0.000001	50	all	70.5 ± 3.1

Table 3: Best per-word tagging accuracies on regularized data

Table 3 shows the best results of the experiment. All improvements are significant.⁸ The parameter values for the best results are divergent. Especially when only the base texts are used, the numbers of Brown clusters yielding the best results differ: For Griseldis, 25 and 50 are the best options; neither are among the three best values for Johannes. However, when using all texts, 50 Brown clusters lead to the second-best results for Johannes and Griseldis. For Johannes, this result is almost as good as the best result. For Griseldis, the difference between the optimal parameters and the combined best result is greater, but it still leads to an increase of about 1% for the tagging accuracy.

Inspecting substituted types and types that are not substituted more closely points toward further steps to improve the system. For instance, Griseldis has at least four variants of ik 'I': ik, yck, yk and ick. Johannes exhibits at least three variants: ik, ic and jk. All of these variants are in the same Brown cluster (50 clusters, all texts). However, when regularizing Griseldis, only ick gets substituted with ic. The reason for this is that string and context similarity are modeled separately, and only the most similar types according to Proxinette can be conflated with types in the same Brown cluster. For yck and yk, there are types that are more similar (according to Proxinette) than one of the actual spelling variants appearing in Johannes. By allowing types other than the most similar types as conflation candidates these variants could be conflated.

Table 4 presents conflated types in Johannes that appear more than 10 times in the text. An 'x' in the fourth column indicates whether the conflations are actually spelling variants. An examination of wrongly conflated types indicates that POS tagging can benefit from these conflations as well. The wrongly conflated types can be divided into four categories:

(1) Morphologically connected types that belong to the same part of speech (derivation or inflection) such as *loue* 'believe, praise, promise' (ISG.PRS.IND (among others)) – *louen* 'believe, praise, promise' (INF (among others)); (2) Types that belong to the same part of speech such as *ioden* 'Jews' – *boden* 'messengers' and *schare* 'cohort' – *hare* 'hair'; (3) Morphologically connected types that belong to different parts of speech (derivation) such as *lose* '(to) loosen' (but may also be: 'replacement, redemption') – *loszheyt* 'flippancy, devilment'; and (4) Types that

⁸The significance was verified using a paired t-test with the tagging results on the 20 parts used for the cross-evaluation. The significance level was set to 0.05 with adjustment to 4 tests using Holm's method.

Freq.	Туре	Conflation	Spelling variant	Translation
11	hochtijt	hochtid	X	'celebration'
12	hijr	hir	Х	'here'
12	scole	schole	Х	'shall'
13	scrift	schrifft	Х	'writing'
14	echter	echtes		'again' (diff. morphology)
15	ghecomen	komen		'come' (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		'cohort'; 'hair'
16	loue	louen		'believe, praise' (diff. inflection)
17	efte	eft	Х	'or'
17	sic	sick	Х	'herself, himself'
18	neman	nemande	Х	'nobody'
21	uadere	vadere	Х	'father'
22	jk	jodoch		'I';'but'
27	comen	komen	Х	'come'
30	lef	leff	Х	'beloved'
43	scolen	scholen	Х	'shall'
61	ioden	boden		'Jews'; 'messengers'
67	uader	vader	Х	'father'

Table 4: Conflations of unknown types in Johannes (2; 0.000001; 50; all)

belong to different parts of speech such as jk 'I' – jodoch 'but'.

For POS tagging, only the third and fourth types are harmful, as the other conflations can still help the tagger to predict the right POS tag. Therefore, for other tasks – e.g. fine-grained POS tagging including morphological tags and lemmatization – the conflation needs to be stricter than for simple POS tagging.

6 Conclusion

In this paper, we have presented a language-independent, unsupervised regularization approach that utilizes string and context similarity and does not make use of any resources other than unlabelled, tokenized texts from the language to be regularized. Applying this approach to POS tagging, we were able to increase the perword accuracy for two historical non-standardized GML texts by about 2% with the optimal parameters and 2% for one text and 1% for the other with parameters giving the combined best result. These are small but still statistically significant improvements.

An analysis of the conflations shows that the algorithm misses some spelling variants because only the most similar types according to Proxinette are considered as conflation candidates. To further improve the algorithm in this direction, we plan to directly integrate the syntactic context into the process of selecting conflation candidates instead of only using it as a filter.

In the study presented in this paper, we divided the data into two parts: one part that defined the target forms and another that was regularized toward these target forms. Such a division comes naturally when using regularization as a preprocessing step for supervised algorithms, as in our evaluation setup. However, for applications that are based on unsupervised algorithms or simple text statistics such as keyword analysis, it would be helpful to avoid such a division. We expect that this would also improve supervised approaches, as the training data would be regularized as well. We did not investigate this in the current study, but our future research will examine this issue.

Additionally, it should be noted that the usefulness of our approach is not limited to POS tagging: All applications that rely on consistent spellings will benefit from regularization. We will explore this in further experiments.

Resources

This paper is created reproducibly using org-mode (http://orgmode.org). The org-files, including all the scripts needed to reproduce the experiments, are available at github (https://github.com/fab-bar/paper-CRH4). This version also includes an appendix with additional data from the experiments.

Acknowledgements

The work of the first and second authors has been funded by the DFG. We would like to thank the anonymous reviewers for their helpful remarks, Sarah Ihden for help with the translations from Middle Low German and Claire Bacher for improving our English. All remaining errors are ours.

Primary data

- Johannes Buxtehuder Evangeliar. GML manuscript from about 1480. Transcribed in the DFG-funded project "Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200-1650)".
- **Griseldis** *Griseldis / Sigismunda und Guiscardus*. GML print of two tales from 1502. Transcribed in the DFG-funded project "Referenzkorpus Mittel-niederdeutsch / Niederrheinisch (1200-1650)".
- **OldbSSP** Oldenburger Bilderhandschrift des Sachsenspiegels. GML manuscript from 1336. Transcribed in the DFG-funded project "Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200-1650)".
- **SaechsWeltchr** *Sächsische Weltchronik.* GML manuscript from the first half of the 14th century. Bremer Hs. der Rezension B (Hs. 16). Transcribed in the

DFG-funded project "Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200-1650)".

References

- Baron, Alistair, Rayson, Paul and Archer, Dawn (2009) Word Frequency and Key Word Statistics in Corpus Linguistics. *Anglistik: International Journal* of English Studies, 20, 41–67.
- [2] Bollmann, Marcel, Dipper, Stefanie, Krasselt, Julia and Petran, Florian (2012) Manual and Semi-automatic Normalization of Historical Spelling – Case Studies from Early New High German. In Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012), LThist 2012 workshop, pp. 342–350, Vienna, Austria.
- [3] Bollmann, Marcel, Petran, Florian and Dipper, Stefanie (2014) Applying Rule-based Normalization to Different Types of Historical Texts – An Evaluation. In Vetulani, Zygmunt and Mariani, Joseph (eds.) *Human Language Technology Challenges for Computer Science and Linguistics*, pp. 166–177, Heidelberg et al.: Springer International Publishing.
- [4] Brown, Peter F., deSouza, Peter V., Mercer, Robert L., Della Pietra, Vincent J. and Lai, Jenifer C. (1992) Class-based N-gram Models of Natural Language. *Computational linguistics*, 18, 467–479.
- [5] Derczynski, Leon, Chester, Sean and Bøgh, Kenneth S. (2015) Tune your Brown Clustering, Please. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2015)*, pp. 110–117, Hissar, Bulgaria.
- [6] Dickinson, Markus and Meurers, Detmar (2003) Detecting Errors in Part-of-Speech Annotation. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03), pp. 107–114, Budapest, Hungary.
- [7] Dipper, Stefanie (2011) Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison. *Journal for Language Technology and Computational Linguistics*, 26, 25–37.
- [8] Dipper, Stefanie, Donhauser, Karin, Klein, Thomas, Linde, Sonja, Müller, Stefan and Wegera, Klaus-Peter (2013) HiTS: Ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics*, 28, 85–137.
- [9] Giménez, Jesús and Màrquez, Lluís (2004) SVMTool: A General POS Tagger Generator based on Support Vector Machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pp. 43–46, Lisbon, Portugal.

- [10] Halácsy, Péter, Kornai, András and Oravecz, Csaba (2007) HunPos: An Open Source Trigram Tagger. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07), pp. 209– 212, Stroudsburg, PA, USA.
- [11] Hathout, Nabil (2009) Acquisition of Morphological Families and Derivational Series from a Machine Readable Dictionary. In Montermini, Fabio, Boyé, Gilles and Tseng, Jesse (eds.) *Selected Proceedings of the 6th Décembrettes*, pp. 166–180, Somerville, MA: Cascadilla Proceedings Project.
- [12] Hathout, Nabil (2014) Phonotactics in Morphological Similarity Metrics. Language Sciences, 46, 71–83.
- [13] Hauser, Andreas W. and Schulz, Klaus U. (2007) Unsupervised Learning of Edit Distance Weights for Retrieving Historical Spelling Variations. In Mihov, Stoyan and Schulz, Klaus U. (eds.) Proceedings of the First Workshop on Finite-State Techniques and Approximate Search, pp. 1–6, Borovets, Bulgaria.
- [14] Kestemont, Mike, Daelemans, Walter and De Pauw, Guy (2010) Weigh your Words – Memory-based Lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25, 287–301.
- [15] Percy Liang (2005) *Semi-supervised Learning for Natural Language*. MEng thesis, Massachusetts Institute of Technology.
- [16] Logačev, Pavel, Goldschmidt, Katrin and Demske, Ulrike (2014) POStagging Historical Corpora: The Case of Early New High German. In Henrich, V., Hinrichs, E., de Kok, D., Osenova, P., and Przepiórkowski, A. (eds.) *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, pp. 103–112, Tübingen, Germany.
- [17] Mitankin, Petar, Gerdjikov, Stefan and Mihov, Stoyan (2014) An Approach to Unsupervised Historical Text Normalisation. In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14), pp. 29–34, New York, NY, USA.
- [18] Naoaki Okazaki (2007) CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs) (URL: http://www.chokkan.org/software/crfsuite).
- [19] Schmid, Helmut and Laws, Florian (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 - Volume 1 (COLING '08), pp. 777–784, Manchester.