Poznan Studies in Contemporary Linguistics

Mining historical texts for diachronic spelling variants

--Manuscript Draft--

Manuscript Number:	PSiCL-D-18-00077R1
Full Title:	Mining historical texts for diachronic spelling variants
Short Title:	Mining historical texts for diachronic variants
Article Type:	Original Study
Section/Category:	Other
Keywords:	spelling variants, OCR, word embeddings
Corresponding Author:	Filip Graliński, Ph.D. Uniwersytet im Adama Mickiewicza w Poznaniu Wydzial Matematyki i Informatyki Poznań, POLAND
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Uniwersytet im Adama Mickiewicza w Poznaniu Wydzial Matematyki i Informatyki
Corresponding Author's Secondary Institution:	
First Author:	Filip Graliński, Ph.D.
First Author Secondary Information:	
Order of Authors:	Filip Graliński, Ph.D.
	Krzysztof Jassem, dr hab.
Order of Authors Secondary Information:	
Manuscript Region of Origin:	POLAND
Abstract:	The paper describes a method for finding diachronic spelling variants in a corpus that consists of historical and modern Polish texts. The procedure applies the Levenshtein distance and the similarity measure determined with a Word2vec model. The method was applied for both words and sub-word units. A sample of spelling variants was manually evaluated and compared against an existing morphological analyser for Polish historical texts. The resulting lists of spelling variants and spelling modernisation rules were used in a text modernisation tool and their contribution was evaluated. The paper also presents an analogous method for finding spelling variants that result from erroneous OCR. The obtained lists of OCR variants and rules may serve for the correction of OCR output.
Suggested Reviewers:	Alexander Koplenig Institut fur Deutsche Sprache
	Michael Piotrowski Universite de Lausanne
	Włodzimierz Gruszczyński Szkole Wyższej Psychologii Społecznej
Opposed Reviewers:	
Response to Reviewers:	

X **1** (1): 1–**??**. Printed in ...

Mining historical texts for diachronic spelling variants

FILIP GRALIŃSKI and KRZYSZTOF JASSEM
Faculty of Mathematics and Computer Science,
Adam Mickiewicz University, Poznań.

(*Received* —)

Abstract

The paper describes a method for finding diachronic spelling variants in a corpus that consists of historical and modern Polish texts. The procedure applies the Levenshtein distance and the similarity measure determined with a Word2vec model. The method was applied for both words and sub-word units. A sample of spelling variants was manually evaluated and compared against an existing morphological analyser for Polish historical texts. The resulting lists of spelling variants and spelling modernisation rules were used in a text modernisation tool and their contribution was evaluated.

The paper also presents an analogous method for finding spelling variants that result from erroneous OCR. The obtained lists of OCR variants and rules may serve for the correction of OCR output.

1 Introduction

Libraries and archives hold large collections of documents in their original paper form, which are accessible only on-site. To enable access to such documents for the public, they are being digitised and stored in digital libraries. The process of digitisation starts with scanning (or photographing). Next comes the conversion of the textual content into digital text, which allows for full-text searching and application of Natural Language Processing (NLP) tools (Piotrowski 2012). The conversion is usually executed in two steps: optical character recognition (OCR) and diachronic normalisation, i.e. transformation of historical spelling into its modern equivalent. Diachronic normalisation may concern: writing of individual words, punctuation, hyphenation or separation of tokens. Here, we focus on spelling variation in individual words.

The main purpose of our work is to facilitate linguistic and historical research through improvements to full-text search engines operating on diachronic corpora, though the data extracted with the methods described in this paper might be useful for NLP for historical texts in general.

2 Diachronic spelling variants

Word spelling variation may occur synchronically and diachronically. The former takes place when more than one spelling of the same word exists in the language in the same period, whereas the latter concerns words whose spellings have changed over time. The distinction between the two types of variations is not straightforward. Usually, word variants co-exist synchronically, and frequencies of their occurrences change over time. This is illustrated in Figure 1 which shows the occurrence frequency of spelling variants: *professor* and *profesor* in Polish texts along the time axis. The historical spelling (*professor*) was observed more frequently in texts before 1865. From then on, the modern spelling has been dominant. We define the date of a diachronic spelling change as the most recent point on the time axis, where the frequency diagrams intersect. Accordingly, we assume that the change *professor* → *profesor* is dated as 1865. The occurrence frequency is defined as an average number of appearances of a word in a million-word sample of a text corpus. The corpus under research, called Odkrywka (Graliński and Wierzchoń 2018), contains 40 billion tokens and consists of Polish publications (mostly newspapers and books) originating mainly from the years 1810–2013.

Diachronic changes may take place due to a general regulation issued by an administrative body – such a change will be referred to as prescriptive. Prescriptive changes are usually but not always mirrored by so called spontaneous changes, observed in real-text corpora. Figure 2 shows the frequencies of spelling variants of the word *puchar* (*goblet*). The prescriptive decision on the change in the spelling of *puhar* took place in 1936 and the real change followed that decision next year, as attested by the diachronic corpus, though was reversed briefly during World War II.

Here, we examine spontaneous spelling changes (no matter whether initiated by a prescriptive change or not). We assume that for each word there exists its dominant spelling for a given year – i.e. most frequently represented in texts. We define diachronic spelling normalisation (or spelling modernisation) as the process of automatic conversion of historical texts into their modern equivalents, in which all words are written according to their presently dominant spelling.

Some spelling changes are more idiosyncratic and are not due to a general rule (such as "*j* after *r* and before a vowel should be changed into *i*"). For instance, the word *aberracja* (*aberration*) was written more frequently with a single *r* in the first half of the 20th century (see Figure 3). The fact that *aberracja* became the standard spelling was not caused by a general and strict rule (actually the general direction of changes is opposite: a double consonant is often replaced with a single one, cf. *terrytorium* → *terytorium*).

This paper describes a method for automatic extraction of pairs of Polish diachronic spelling variants. This is a step towards improving the quality of spelling modernisation of Polish texts: firstly, the list may serve for lexicon-based modernisation, secondly, the pairs help induce modernisation rules for rule-based modernisation. The approach suggested here uses the Levenshtein distance[1] as one of the

[1] *Levenshtein distance* between two words is the minimum number of basic editing op-

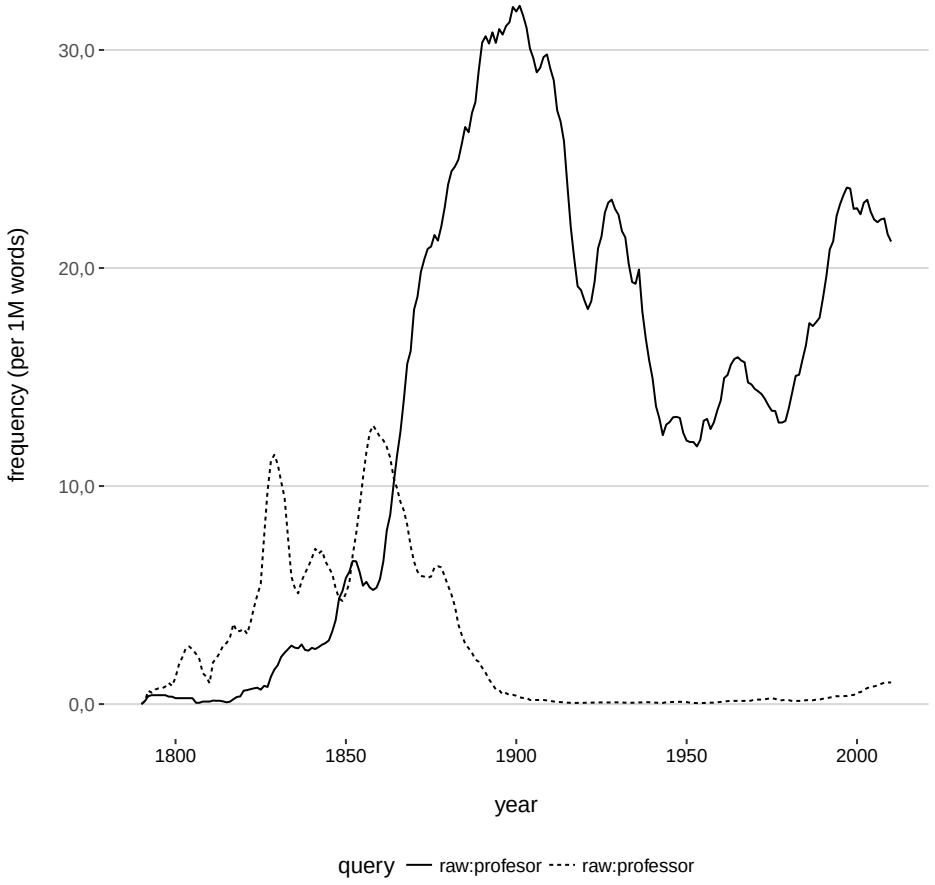

Fig. 1. The frequency of *profesor* and *professor* in the Odkrywka corpus.

criteria for identifying potential pairs. In (Bollman *et al.* 2011) the authors align a historical and a modern version of the Luther Bible to automatically derive text modernisation rules. After the text sources are aligned on the sentence and word level, the Levenshtein distance algorithm searches for aligned words that differ by a few edit operations, and stores the types of operations used in transforming one word into another. Here, we operate on a non-aligned corpus of large volume (of both historical and modern texts). Therefore, to decrease the search space we limit the Levenshtein distance to one edit. Additionally, we use Word2vec word embedding as another criterion for restricting the set of candidates.

Historical spellings are considered in the OCR post-correction based on the so-

erations (deletions, insertions, substitutions) that is necessary to change one word into another. For instance, the Levenshtein distance between the words *three* and *trees* is 2, since two editing operations are needed to turn the former into the latter: *h* must be deleted and *s* — inserted.

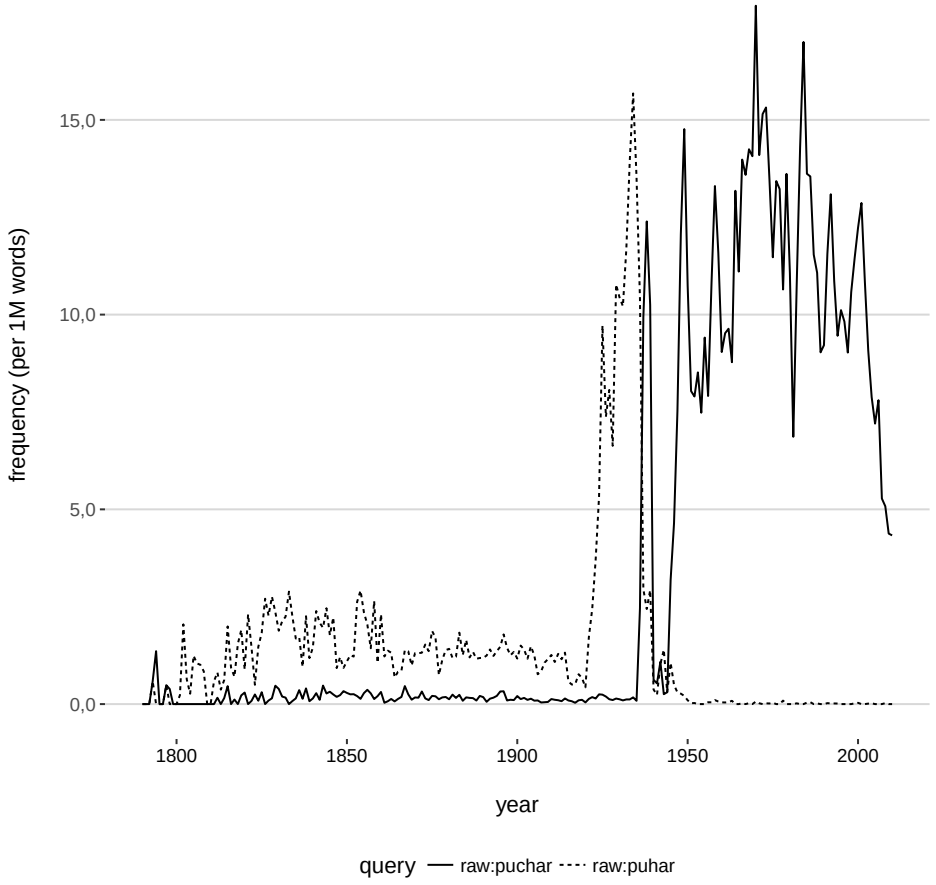

Fig. 2. The frequency of *puhar* and *puchar* in the Odkrywka corpus.

called profiling of OCR-ed historical texts (Reffle and Ringlstetter 2013) (Fink *et al.* 2017). No Word2vec model has been used there.

3 Word2vec model

Word embeddings are numerical representations of words, which capture their meaning and semantic relationships by examining contexts they are used in. The concept of the so-called Word2vec word embedding was introduced by Mikolov *et al.* (2013) The authors apply the Skip-gram language model (used to predict what words are most likely to occur in the left and right context of a given word) to find vector word representations. Mikolov *et al.* claimed that:

> [...] simple vector addition can often produce meaningful results. For example, w2v(*Russia*) + w2v(*river*) is close to w2v(*Volga River*), and w2v(*Germany*) + w2v(*capital*) is close to w2v(*Berlin*) [w2v denotes Word2vec vector representation]

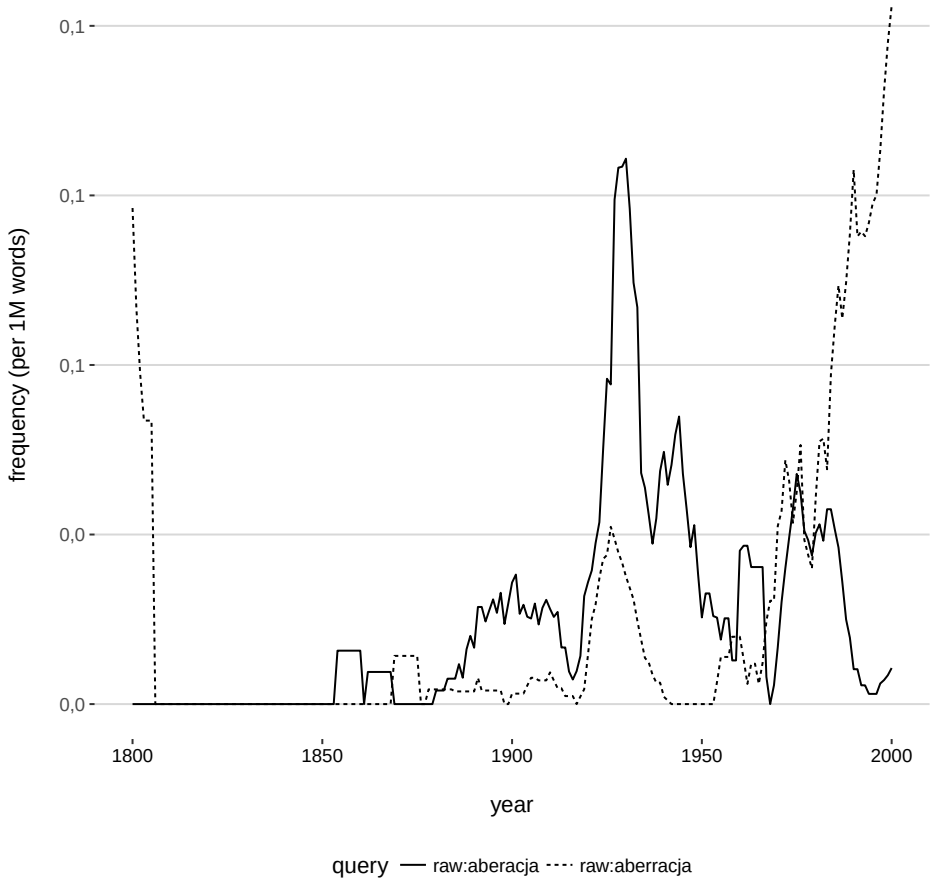

Fig. 3. The frequency of *aberacja* and *aberracja* in the Odkrywka corpus.

The idea of word embeddings is rooted in *distributional hypothesis* by Zellig Harris (Zellig 1954): words with similar distributions (contexts) have similar meanings. John Firth famously summarised the idea as: "You shall know a word by the company it keeps!" (Firth 1957). Technically, Word2vec embeddings are obtained via a stochastic process of iteratively improving vectors (lists of numbers) assigned to words in such a way that words tending to occur in similar contexts would be located in close distance to each other a multidimensional space.

Word embeddings have been successfully applied to various NLP tasks, such as machine translation, named entity recognition, part-of-speech tagging, or language modelling. Here, we suggest another application: finding diachronic spelling variants.

4 Searching for diachronic spelling variants

To obtain a large list of diachronic spelling variations for Polish we have developed a procedure that searches for pairs of diachronically equivalent words in a large text corpus. The procedure applies the Word2vec model and is executed semi-automatically on the Odkrywka text corpus. As Odkrywka consists mainly of digitised versions of scanned documents, the procedure is supposed to handle texts that contain a significant number of OCR errors.

We assume that edit types are represented by the notation $\alpha \rightarrow \beta$ (*alpha* is substituted by β), e.g. the edit type between spelling variants *hiperprodukcya* and *hiperprodukcja* is represented by $y \rightarrow j$ (cf. *historical patterns* in OCR profiling (Fink *et al.* 2017)). We assume, however, that neither α nor β can be empty (such edit types would not be informative enough when being manually checked), therefore when a letter is deleted or added, two edits — one with the letter to the left and one with the letter to the right — are considered and the one which is more frequent across all variant pairs is finally selected. For instance, for the spelling variant *intellektualny* of the adjective *intelektualny* (*intellectual*) we consider two edits ($ll \rightarrow l$ and $le \rightarrow e$) and we choose the more frequent one ($ll \rightarrow l$, which occurred in 80 pairs, whereas $le \rightarrow e$ – in no other pair).

The mining procedure consists of the following steps:

1. Create a Word2vec model for the diachronic corpus (use only unannotated plain text, no metadata is considered here).
2. Take a list of lemmata from a modern dictionary of inflected forms.
3. Filter out short lemmata (shorter than some threshold λ) – as short words tend to introduce a significant amount of "noise" (admittedly, some "interesting" words are lost here, as frequent words tend to be short; still, the problem is that many short words are infrequent and most of their occurrences in digitised texts are just irregular OCR "hallucinations").
4. Manually prepare a regular expression ρ to filter out words likely to be OCR misrecognitions (of course not all OCR mistakes will be detected this way, but at least some simple cases will be filtered out).
5. Manually prepare the edit blacklist B (list of changes which are typical for OCR mistakes and/or are highly unlikely to represent real changes in spelling conventions).
6. For each lemma w represented in the Word2vec model:
   (a) find all words v represented in the Word2vec model such that the cosine similarity between w and v exceeds some threshold σ *and* the Levenshtein distance between w and v is smaller than or equal to δ (i.e. we want semantically similar words which are spelled in a similar way),
   (b) for each v: discard v if it is shorter than λ,
   (c) discard v if it is already present in the dictionary of inflected forms (as an inflected form, not necessarily as a lemma),
   (d) discard v if it matches the regular expression ρ,
   (e) determine the edits between v and w,
   (f) discard v if one of the edits is in B,

(g) otherwise return (v, w).

In other words, for a given lemma w we assume that its potential diachronic spelling variant v:

- is similar to w according to the Word2vec model,
- is close to w orthographically,
- does not belong to a modern vocabulary
- and is unlikely to result from erroneous OCR processing.

This procedure was applied to the Odkrywka text corpus with the following settings:

- PoliMorf (Woliński *et al.* 2012) was used as a source of modern lemmata and inflected forms,
- Word2vec was instructed to consider only words whose frequency is greater than or equal to 6 (rarer words were disregarded),
- λ was set to 6, i.e. only words composed of at least 6 characters are considered,
- the regular expression ρ was `[0-9.ñø]|^.?-|-.?$` (discard strings with digits, dots, unusual characters anywhere, or with a hyphen in an unusual position, i.e. at or near the end or beginning of the word),
- σ was set to 0.74 (the value was tuned manually during the initial stage of the experiments),
- δ was set to 1, i.e. only one simple change is acceptable (a single character is inserted, a single character is deleted or a single character is replaced by another character),
- some specific digraphs (*ch*, *cz*, *dz*, *dź*, *dż*, *gz*, *ks*, *rz*, *sz*) were treated as if they were single letters, as they are likely to be substituted with a single letter (e.g. *x* for *gz* and *ks*, *rz* for *ż*) in such a way that the simple Levenshtein distance would have been equal to 2.

For computational efficiency (in order to shorten processing time) it was also assumed that the potential variants should not differ in the first character.[2] In order to account for some likely exceptions to this assumption, *i*, *j* and *y* as well as *g*, *k* and *x* were treated as one — cf. diachronic variants *iakoby*/*yakoby* for *jakoby* (*as if*) or *xiążka* for *książka* (*book*).

Some edit types are inherently hard to classify as they could arise roughly equally likely due to spelling change as due to OCR misrecognition. For example, $j \rightarrow i$ is a very frequent diachronic variation (*Marja* $\rightarrow$ *Maria*), but it could also occur as a result of the letter *j* misrecognised as *i* (e.g. as a result of the printing ink damaged at the bottom of the letter, as exemplified by the spurious word *angielskj*). Few frequent edit types like this were covered with an extra condition encoded as a hand-crafted regular expression, for instance $j \rightarrow i$ is treated as a diachronic spelling variant on condition that *j* is preceded or followed by a vowel.

[2] Using techniques (including efficient look-up) implemented recently in the Magnitude Python library (`https://github.com/plasticityai/magnitude`) would probably obviate the need for such a limitation.

This procedure yielded 42,808 pairs representing 1,796 edit types. The edit types which occurred in at least 5 pairs were manually inspected and 71 edit types were accepted as related to changes in spelling conventions (the remaining edit types were added to the blacklist B). The most frequent types of edits are presented in Table 1. After using the blacklist, we obtained a list of 5,729 pairs representing diachronic spelling variants.

#	edit	variant	word
1404	j → i	salustjusz	salustiusz
862	y → j	oyrzyński	ojrzyński
736	i → j	wydainość	wydajność
463	y → i	dyablica	diablica
166	o- → o	głucho-niemy	głuchoniemy
130	ij → i	kaligrafija	kaligrafia
110	i → y	gabardina	gabardyna
107	é → e	słyszéć	słyszeć
101	ss → s	essencja	esencja
100	z → s	senzacje	sensacje
92	yj → j	konwencyja	konwencja
91	s → z	gisela	gizela
87	x → ks	kalixta	kaliksta
73	ie → e	gieneza	geneza
62	y → e	zależyć	zależeć
57	ó → u	jakóbowice	jakubowice
47	l → ll	holywood	hollywood
44	n → nn	lineusz	linneusz
43	v → w	cromvell	cromwell
40	e- → e	gdzie-niegdzie	gdzieniegdzie
39	c → k	fiasco	fiasko
35	dz → c	przemódz	przemóc
32	sz → s	szkwara	skwara
32	ch → h	scheridan	sheridan
30	u → ó	struża	stróża
30	tt → t	attakować	atakować
30	nn → n	tennisista	tenisista
29	w → v	harward	harvard
29	gs → s	malborgski	malborski
28	s → ss	danielson	danielsson
27	th → t	ementhaler	mentaler
27	ws → s	królewstwo	królestwo
26	rr → r	rrezydent	rezydent
26	r → rr	heriot	herriot
25	y → o	chrystowska	chrostowska
25	k → c	marakaibo	maracaibo
23	pp → p	hippurowy	hipurowy
22	t → d	kronsztat	kronsztad
22	mm → m	programmat	programat
21	s → ws	zmartwychstanie	zmartwychwstanie

Table 1. The most frequent types of edits related to spelling variants

The procedure for mining spelling variants is based on two conditions: the two words need to be close in terms of both Levenshtein distance *and* a Word2vec model. Why not simply use the editing distance? This was tried at first, but the number of false positives was too large — when a lot of words are considered, a great amount of spurious pairs were generated (especially for names, where it is quite likely to obtain a completely non-related, let us say, surname by changing just one letter).

Simply both surface and semantic (contextual) similarity is necessary. Note that more advanced vector models which take into consideration the characters from which a word is composed (e.g. fastText which composes a word embedding out of character n-gram embeddings) might be less effective as the semantic and surface form are not clearly separated in them.

4.1 Inflected forms

Only base forms of common nouns were considered in the procedure described so far. Inflected forms of spelling variants were generated separately by analogy. For instance, *ambassada* was identified as a diachronic variant of the noun *ambasada* (*embassy*), which has the following inflected forms: *ambasady*, *ambasadzie*, *ambasadę*, *ambasadą*, *ambasado*, *ambasad*, *ambasadom*, *ambasadami*, *ambasadach*. Consequently, the following pairs of diachronic variants were generated: *ambassady* → *ambasady*, *ambassadzie* → *ambasadzie*, *ambassadę* → *ambasadę*, *ambassadą* → *ambasadą*, *ambassado* → *ambasado*, *ambassad* → *ambasad*, *ambassadom* → *ambasadom*, *ambassadami* → *ambasadami*, *ambassadach* → *ambasadach*. Note that when the ending of the base form is modified, only some or none of the inflected forms can be turned into spelling variants, e.g. for the variant *pobiedz* of *pobiec* (*run*) no inflected forms will be generated as the ending *-c* is present only in the base form.

When their inflected forms were generated, the list of diachronic spelling variants expanded to 104,064 forms.

4.2 Variants with more than one edit

A limitation of this method of mining spelling variants is that only variants differing by exactly one modification are considered, whereas there exist examples of spelling variants which differ by more than one edit. For instance, the word *komisja* (*commission*) not only has such diachronic variants attested by the corpus as *komisyja*, *komisija*, *komissja*, *kommisja*, but also *kommisya*, *komissya* (the Levenshtein editing distance to the modern form is 2) or even *kommissya* (of Levenshtein distance 3). The problem might be alleviated to some extent by first applying general rules based on regular expressions (e.g. the one changing the suffix *ya* into *ja* after the letter *c*, *s* or *z*). Nonetheless, some irregular variants (such as *kommissja*) would be still unrecoverable. In order to fully account for more "distant" spelling variants (the ones of Levenshtein distance 2), the whole procedure was iterated once more, now with spelling variants obtained in the first round used as "target" forms (σ was set to a higher value, 0.8, in order to avoid "drifting" to a completely different meaning). This way, 461 additional spelling variants were obtained (7,127 inflected forms). Then the procedure was repeated to gather spelling variants of Levenshtein distance 3 and 19 more variants were obtained. Spelling variants of Levenshtein distance 4 are unlikely (and false positives would start to appear), so the procedure was not continued.

Interestingly, the enhanced procedure yielded valid diachronic spelling variants

differing in *one* edit, e.g. the variant *sekretaryat* of *sekretariat* (*secretariat*) was found as the result of two edits $y \rightarrow j$ and $j \rightarrow i$. What was the reason why the direct edit $y \rightarrow i$ was not found when the main procedure was applied? The edit $y \rightarrow i$ *was*, of course, tagged manually as representing a spelling change, but the Word2vec similarity between *sekretaryat* and *sekretariat* was just below the threshold ($0.733 < \sigma = 0.74$), whereas the "roundabout" similarities between *sekretaryat* and *sekretarjat* as well as *sekretarjat* and *sekretariat* are above the higher threshold of 0.8 (respectively, 0.882 and 0.877). This may be slightly surprising, but it actually reflects the history of Polish spelling system — the original spelling was *-rya-*, then it was changed into *-rja-* and finally into *-ria-* during the 1936 prescriptive spelling change (see Figure 4). There are two possible reasons why *sekretaryat* and *sekretariat* were not similar enough for Word2vec. First, the contexts in which the two variants occurred in, respectively, 19th and 20th century might differ too much (independently of the change in its surface form). Second, the Word2vec model *does* try to predict the way the words in the context are spelled (even if syntactic and semantic features are more important) — in this sense, the contexts in which *sekretaryat* and, for instance, *sekretarjat* appear are more similar than the contexts for *sekretaryat* and *sekretariat*.

5 Diachronic vs OCR-generated variants

Note that if we *accepted* variants with edits in the blacklist B rather than filter them out, we would obtain a list of frequent OCR errors, which would be useful for OCR post-correction. Actually, one may ask why distinguish OCR mistakes and spelling variants at all, as for natural language processing it does matter which is which (all that is wanted is text in the cleanest form possible, written using modern conventions) and when a search engine for historical research is concerned, a user (e.g. a historian) wants them all normalised anyway. The answer is that:

- in some types of linguistic research (e.g. when spelling evolution is analysed) diachronic spelling variants should be distinguished (whereas OCR mistakes are rarely of any direct interest to a linguist),
- in full-text search both OCR errors and spelling variants should be considered (e.g. query *demokracja* should match both *demokracya* and what was misrecognised as *demoknacja* due to the letter *r* OCR-ed as *n*), but the text shown in search results (e.g. in titles and snippets), should be given in the original form (with fixed OCR errors, but in the original spelling),
- we need to know the original spelling to mark a given word when it is presented in a "clipping" according to the principles of *photodocumentation* (Wierzchoń 2010) — but OCR errors should, again, be fixed if possible.

In view of this, B was actually split into two subsets: one for OCR errors and one for all unclear cases (the edit itself is not enough to decide, the edit represents a synchronic variant, etc.). The misrecognitions obtained with the edits from the former group were added to the OCR cleaning procedures. Also, inflected forms were generated in the same way as the inflected forms of diachronic spelling variants

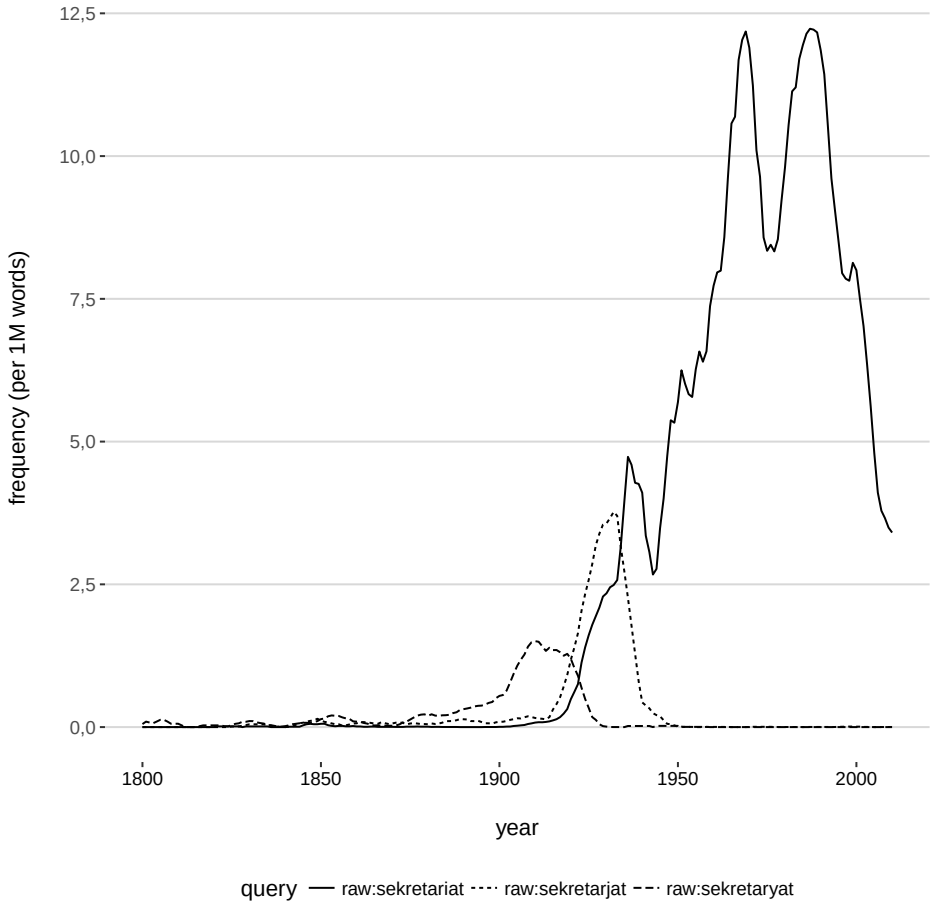

Fig. 4. The frequency of *sekretaryat*, *sekretarjat* and *sekretariat* in the Odkrywka corpus.

were obtained (see Section 4.1); misrecognitions with the initial letter upper-cased are added as well unless such a word is homonymous with a known proper name. The number of edits tagged as likely to be the result of OCR errors was 250. They yielded 27,261 pairs (of an OCR error and its correction) and 466,433 pairs when inflected forms were generated. The most frequent types of edits related to OCR errors are presented in Table 2.

5.1 Handling surnames

Proper names were also considered in this procedure. (Even though the Word2vec model was based on a lower-cased text, when a spelling variant or an OCR mistake is identified, the first letter of its target word is upper-cased and verified in the lexicon of inflected forms.) It was, however, noticed that a significant number of non-related pairs of surnames were returned by the procedure. Simply, a surname might be close

#	edit	OCR misrecognition	correct word
1808	l → i	spodnl	spodni
609	z → ż	rewanz	rewanż
579	a → o	tarnawiecki	tarnowiecki
454	o → e	poomat	poemat
422	é → ć	nalepié	nalepić
417	s → ś	kusmierczyk	kuśmierczyk
362	e → ę	postep	postęp
330	a → ą	zajmujacy	zajmujący
293	i → ż	połoienie	położenie
289	e → ć	rozwijae	rozwijać
280	ó → ć	włożyó	włożyć
252	u → o	wujewoda	wojewoda
241	t → ż	podrótnik	podróżnik
210	e → o	tewarzystwo	towarzystwo
197	o → a	łowiński	ławiński
195	i → ł	przybyia	przybyła
185	ź → ż	odróźnić	odróżnić
179	c → o	spcsób	sposób
170	n → rz	pnegląd	przegląd
164	o- → o	disco-polo	discopolo
140	a → u	juliasz	juliusz
137	o → n	koogres	kongres
129	u → a	pustuła	pastuła
127	l → r	polwać	porwać
126	a → s	miatrzostwo	mistrzostwo
125	t → i	stosunkt	stosunki
117	z → ź	woznicki	woźnicki
109	t → ć	dostosowat	dostosować
93	l → j	kuracla	kuracja
92	b → s	łabiński	łasiński
88	j → y	politjka	polityka
87	l → f	józela	józefa
87	e → ie	legerski	legierski
83	ę → e	aleksandęr	aleksander
83	j → ł	mjodzież	młodzież
79	i → ie	odpowiedzić	odpowiedzieć
78	i → ź	zakainie	zakaźnie
78	t → l	ptaszczyk	plaszczyk
76	d → n	tardowa	tarnowa
72	i → f	korianty	korfanty
72	i → a	licyticja	licytacja

Table 2. The most frequent types of edits related to OCR errors

to an unrelated surname in a Word2vec model. The reason is that all surnames are contextually similar to one another and are distinct from other words. For instance, surnames *Narzyński* and *Nurzyński* are not related to each other, but as they differ in one letter and are similar according to the Word2vec model, *Narzyński* would be returned as an OCR misrecognition of *Nurzyński*. Actually, a number of extracted edit types was neither due to diachronic change or OCR mistakes but they were random replacements in unrelated surnames, e.g. $d \rightarrow g$ for pairs of surnames such as *Rodacki*/*Rogacki*, *Rydlewicz*/*Ryglewicz*, *Macuda*/*Macuga* (the surnames are not related diachronically or genetically).

Also, in some cases a pair of surnames might be genetically related, but there is no guarantee they were used by the same person (or even the same family), i.e. whether the spelling of a surname was changed in accordance with general

changes in orthography and whether it should be normalised. Consider for instance surnames *Finkelstein*, *Finkelsztain*, *Finkelsztajn*, *Finkelsztein*, *Finkelsztejn*, all likely emerging from a common origin, possibly twisted between German, Polish, Russian and Yiddish and nowadays used as separate surnames (i.e. they should not be covered by diachronic spelling normalisation). What is more, outdated spelling rules have been petrified in some surnames, e.g. the *modern* surname *Gieysztor* (according to the 1936 spelling reform it should have been changed to *Giejsztor*).

In order to avoid treating such forms as spelling variants or OCR mistakes, the words were also filtered out using an extensive list of nearly 400,000 Polish surnames.[3] To be precise, a pair of words was removed if both words were found there (after the initial letters were upper-cased). The filter turned out to be too strict, as there was a significant number of valid spelling variants coinciding with rare surnames, e.g. *biórko* is more likely to be a diachronic variant of *biurko* (*desktop*) rather than an unusual surname *Biórko* lower-cased. That is why it was decided that the surnames coinciding (when lower-cased) with a word listed in the PoliMorf lexicon should be exempted from the filter.

6 Mining sub-word diachronic variants

So far we have assumed that only the whole words are considered when mining for diachronic variants and the extracted diachronic variants are used via a simple dictionary look-up when diachronic normalisation is performed. This improves diachronic normalisation significantly, but some special cases are still left out (even for frequent edit types).

1. Inflected forms were covered, but not those derivatives which are not frequent enough to be taken into account by the procedure, e.g. the procedure returned spelling variants *hypnotyczny*, *hypnotyzer*, *hypnotyzm*, *hypnotyzować*, *zahypnotyzować* (as well as their inflected forms) for words, respectively, *hipnotyczny* (*hypnotic*), *hipnotyzer* (*hypnotist*), *hipnotyzm* (*hypnotism*), *hipnotyzować* (*hypnotise*), *zahipnotyzować* (*hypnotise*), but not *odhypnotyzować* (*dehypnotise*), *hypnotyzerski* (*hypnotist's*), *posthypnotyczny* (*posthypnotic*) (actually the variant *hypnoza* of the base word *hipnoza* (*hypnosis*) was not retrieved either, as the semantic distance between *hipnoza* and *hypnoza* was slightly below the threshold — probably due to the fact that hypnosis had different cultural contexts in the 19th and 20th century); the procedure is simply unable to learn that the substring *hypno* should be always modernised into *hipno*.
2. If the variant is "damaged" by an OCR error in some other part of the word (other than where spelling change occurred), the word would not be modernised by the procedure described so far. For instance, one might expect the word *hypnotycznyoh* (*hypnotycznych* in which the letter *c* was misrecognised

[3] `http://www.futrega.org/etc/nazwiska.zip`

as *o*) will be normalised to *hipnotycznyoh*, but this will not happen (unless the OCR post-correction is done first).

3. The procedure assumes that the modern form is listed in the (modern) lexicon of inflected forms. Unfortunately, this means that archaic words which would have undergone spelling change if they had been still in use will not be recovered.

In order to, at least partially, alleviate this issue, an attempt was made to find *sub-word spelling variants*, i.e. character n-grams which are distinctive enough to be safely substituted with their modern counterparts within any word. For instance, *hypno* could be substituted by *hipno* as no modern Polish word has the former prefix/infix.

Initial experiments suggested that, for texts with a high level of OCR noise, a rather conservative setup needs to be assumed (otherwise, OCR mistakes as well as foreign names and words would be spuriously modified), i.e. a character n-gram is accepted as a sub-word spelling variant if:

- it occurs in at least 3 diachronic spelling variants (e.g. as was mentioned, *hypno* occurred in 5 words mined),
- it consists of at least 5 characters,
- it does not occur in *any* word listed in the lexicon of modern inflected forms,
- the edit does not consist in changing the letter *i* into some other letter (this was hard-coded as it turned out that *i* is quite often mistaken for some other letter, e.g. *ż* or *l*, misrecognised by OCR),
- it does not contain any shorter sub-word spelling variant (this condition is just for technical reasons — to prune redundant sub-word units and speed up processing time).

The above set-up resulted in 654 sub-word spelling variants. When sub-word spelling variants were applied, there was an improvement, albeit small, in the test results (the parameters of the mining procedure were tuned on a separate development set). (The improvement would likely be more pronounced on data without OCR noise.) Interesting and significant patterns were extracted, for instance the diachronic change *burgs* → *burs* was identified (exemplified by adjectives referring to cities, e.g. *hamburgski* → *hamburski*, *petersburgski* → *petersburski*).

As an interesting by-product of work on sub-word diachronic variants, a number of misspelled words were identified in the PoliMorf lexicon: *hypnotyzerka*, should be *hipnotyzerka* (*female hypnotist*), *hyperbola*, should be *hiperbola* (*hyperbole*), *hypertekst*, should be *hipertekst* (*hypertext*). Actually, the mistakes were made in the Polish lexicon for Hunspell[4] spell-checker and then carried over to PoliMorf (Woliński *et al.* 2012) and PoliMorfologik[5] lexicons of Polish inflected forms.

[4] `https://hunspell.github.io/`, retrieved 1 December 2018
[5] `https://github.com/morfologik/polimorfologik`, retrieved 1 December 2018

7 Mining specific diachronic spelling variants

The methods described so far are capable of extracting only restricted types of diachronic variants:

- the variant must represent a frequent edit type (i.e. it must be a rather regular spelling change); consequently, idiosyncratic spelling changes, exemplified by just a few words (or even by just one) will not be covered (note that they may actually be frequent words);
- only base form (lemma) is considered; the inflected forms of an old spelling variant are automatically generated, but if a spelling change is visible only in an inflected form (but not in the lemma), the variant will not be extracted.

In order to find atypical diachronic variants of frequent words (including inflected forms), another round of text mining was carried out:

1. The whole text corpus was reprocessed using the diachronic normalisation procedure with the words and character n-grams mined so far *and* OCR post-correction was done using the list of OCR mistakes extracted as described in Section 5.
2. A new Word2vec model was generated using cleaner textual data. This new model could be expected to be more reliable. What is more, some words might start to appear when their old variants or OCR misrecognitions are cleaned up (as they might exceed the threshold for the number of occurrences set when generating the Word2vec model).
3. Again we looked for pairs of words close both in terms of Levenshtein distance (i.e. its written form) and Word2vec similarity (i.e. semantic similarity) for which one word is confirmed by the modern lexicon of inflected forms and another word — is not. This time, however, we are only looking at words in isolation, not edit types. Furthermore, we do not require words be in base forms and no lower limit for word lengths is used.
4. Words (the ones written with a non-standard spelling) are sorted by their frequency and then they are manually checked starting from the most frequent ones and are tagged as being a diachronic spelling variant, an OCR mistake or as a coincidental pair.

The top 4,000 pairs were checked manually, out of which 94 were classified as diachronic variants (and almost all the rest were OCR mistakes). Most of them were generated by the $i \rightarrow j$ and $é \rightarrow e$ transformations, but there were some unique variants discovered, not obvious for a modern Polish native speaker, e.g.: *Rossja* → *Rosja* (*Russia*), *twoję* → *twoją* (the feminine accusative of *twój* (*your*)), *cóś* → *coś* (*something*), *armija* → *armia* (*army*), *bużet* → *budżet* (*budget*).

8 Manual evaluation of diachronic variants

In order to assess the quality of diachronic variants obtained with all the procedures described herein, a random sample of 500 spelling variants (including the inflected and upper-cased forms) was manually inspected:

- it was checked whether the word is really a diachronic spelling variant (not an OCR error; scans of historical publications were consulted manually if needed);
- if this was the case, the word was tested against the morphological analyser created for Polish texts from the years 1830–1913 by Derwojedowa *et al.* (2014), see `http://sgjp.pl/morfeusz/demo-f19`.

It turned out that 434 (86.8%) words were real diachronic spelling variants, out of which only 109 (26.1%) were recognised by the morphological analyser.[6] (The correctness of the analysis was not checked, only the fact that a tag other than `ign`, meaning an unknown word, was returned; the 16 words with hyphens were not counted here, as the analyser treats such words as compounds.) This means that the inexpensive semi-automatic procedure described here was able to yield a significant number of forms that were not recovered with methods used by Derwojedowa *et al.* A sample of such words is given in Table 3. Some of them are rather theoretical, e.g. *dyferencyo* → *dyferencjo*, the vocative form of *dyferencya* → *dyferencja* (*difference*), but the base form *dyferencya* is not recognised by the analyser, anyway. On the other hand, some spelling variants might be older than the lower bound for the time range assumed in the project under which the analyser was developed (1830) — though not much older as there are not many texts earlier than late 18th century in the Odkrywka corpus, on which the Word2vec model was trained.

9 Comparison and evaluation against related work

Applying diachronic spelling variants for the improvement of a text modernisation tool was implemented in the VARD tool (Rayson *et al.* 2005). There, a manually collected lexicon of variant pairs served to modernise texts from Early Modern English. Jurish (2010), in turn, applied probabilistic methods for "canonicalisation" of German words, taking into account their context (without, however, Word2vec-like embeddings). Gotscharek *et al.* (2011) designed a lexicon for historical German texts, by comparing vocabulary from historical German. Bollman *et al.* (2011) extracted context-dependent rules for the modernisation of Early German texts. Their approach resulted in a satisfying accuracy (91%) when run on the text the rules were trained on (i.e. Luther Bible). Much lower performance was reported when the tool was applied to a different text collection originating from the same period (ca 42%).

It is believed that simple accuracy is not the right choice of an evaluation metric for diachronic normalisation (Jassem *et al.* 2017), as the evaluation of text normalisation, in which most of the input text is expected to remain unchanged, should be based on precision and recall – expected changes should be rewarded, unwanted modifications should be penalised and copied input symbols should not affect the score. The metric introduced by Jassem *et al.* (2017), CharMatch, is loosely based

[6] The on-line version of the analyser available on April, 26th, 2018 was used.

old spelling	new spelling
bigoterjom	*bigoteriom*
dyferencyo	*dyferencjo*
Dyluwjalnymi	*Dyluwialnymi*
Dyzlokacyom	*Dyslokacjom*
Felixą	*Feliksą*
hjacynt	*hiacynt*
iałmużnami	*jałmużnami*
imperjalistycznej	*imperialistycznej*
komercyalni	*komercjalni*
Kurjalnemu	*Kurialnemu*
Legjonową	*Legionową*
Mahmut	*Mahmud*
Muzykologja	*Muzykologia*
Oblegaiącego	*Oblegającego*
przyzwyczaiasz	*przyzwyczajasz*
sanitaryuszka	*sanitariuszka*
senzacyami	*sensacjami*
Supplikowali	*Suplikowali*
terjerów	*terierów*
territorjum	*terytorium*

Table 3. A sample of old spelling variants not recognised by the morphological analyser

on MaxMatch (Dahlmeier and Ng 2012), adapted for character edits (rather than word edits) and follows such assumptions.

A solution prepared earlier, based on hand-crafted rules and lexicons {b73859}[7] yielded CharMatch = 0.4656 on a test set of 150 historical texts (each composed of 500 words). (The test set is available at `git://gonito.net/challenge/dia-norm` and was prepared along the lines described in (Jassem *et al.* 2017)). With all the data mined using methods described in this paper {99db90} the result rose to 0.6093.

10 Conclusions

The paper reports on experiments that aim at compiling a list of diachronic spelling variants for Polish texts. As compared to previous attempts carried out for English and German, our approach differs in two aspects: it does not require historical texts to be aligned to modern equivalents (which results in much higher volume of the corpus), and it applies Word2vec word embedding as the similarity criterion

[7] This is the reference code to a repository stored at Gonito.net. The repository may be also accessed by going to `http://gonito.net/q` and entering the code there.

for spelling variants. The results of experiments, i.e. the list of variant pairs and the set of context-dependent spelling rules derived from them, help significantly improve the text modernisation tool for the Polish language. We also show that similar methods may improve correction of OCR errors in digitised texts.

Not only can the methods described here improve processing historical texts, but can also give purely linguistic insights into the evolution of the Polish spelling system (or any other language the spelling of which underwent a significant change).

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*, pp. 34–42.

Dahlmeier, D. and Ng, H. T. (2012) Better Evaluation for Grammatical Error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 568–572. Association for Computational Linguistics.

Derwojedowa, M., Kieraś, W., Skowrońska, D. and Wołosz, R. (2014) Zasób leksykalny polszczyzny II poł. XIX wieku a możliwość automatycznej analizy morfologicznej tekstów z tego okresu. In *Leksyka języków słowiańskich w badaniach synchronicznych i diachronicznych*. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, Toruń.

Fink, F., Schulz, K. U., and Springmann, U. (2017) Profiling of OCR'ed Historical Texts Revisited. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*. ACM. 2017, pp. 61–66.

Firth, J. R. (1957) A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*. Blackwell, Oxford, p. 11.

Gotscharek, A., Reffle, U., Ringlstetter, Ch., 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 (IJDAR)*, **14(2)**:159–171.

Graliński, F. and Wierzchoń, P. (2018) Odkrywka, czyli leksykografia diachroniczna *live Prace Filologiczne* (in press).

Jassem, K., Graliński, F. and Obrębski, T. (2017) Pros and Cons of Normalizing Text with Thrax. In *Proceedings of the 8th Language and Technology Conference*, pp. 230-235.

Jurish, B. (2010) More than words: using token context to improve canonicalization of historical German. *Journal for Language Technology and Computational Linguistics*, **25.1**:23–39.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient estimation of word representations in vector space. *CoRR*, **abs/1301.3781**.

Piotrowski, M. (2012) *Natural language processing for historical texts*. Morgan & Claypool Publishers.

Rayson, P., Archer, D. and Smith, N. (2005) VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. *Proceedings of Corpus Linguistics 2005*.

Reffle, U. and Ringlstetter, Ch. (2013) Unsupervised profiling of OCRed historical documents. *Pattern Recognition*, **46.5**: 1346–1357.

Wierzchoń, P. (2010) Torując drogę teorii lingwochronologizacji. *Investigationes Linguisticae* **XX**:105–185.

Woliński, M., Miłkowski, M., Ogrodniczuk, M. Przepiórkowski, A. and Szałkiewicz, Ł. (2012) PoliMorf: a (not so) new open morphological dictionary for Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pp. 860–864, Istanbul, Turkey.

Zellig, S. H. (1954) Distributional structure. *Word*, **10.2-3**:146–162.