– Invited Talk –

Variance in Historical Data: How bad is it and how can we profit from it for historical linguistics?

Stefanie Dipper
Linguistics Department
Ruhr-Universität Bochum
dipper@linguistics.rub.de

The most striking feature of historical language data is probably the amount of variance, in particular variance of spelling. For example, in a Bavarian manuscript from the 16th century, written by one author, we find eight different spellings of the word *Kreuz* 'cross':

> *creuecz*, *cruecz*, *kreevcz*, *kreucz*, *kreuecz*, *krevcz*, *krevecz*, *kruecz*.

If we look at the Anselm corpus 1 , which contains about 50 manuscripts and prints from different dialects of Early New High German[1] (1350–1650), there are in total 50 different spellings of that word:

> *chraewcz*, *chrawcz*, *chrawecz*, *chreitz*, *chreucz*, *chreuecz*, *chreutz*, *chrevcz*, *chrevtz*, *chrewcz*, *chrewczt*, *chrewecz*, *chrewtz*, *chrvtz*, *creucz*, *crecz*, *creuecz*, *creutz*, *cretz*, *crewcz*, *crewtz*, *crucz*, *cruecz*, *cruetz*, *cruicz*, *cruitz*, *cruiz*, *crutz*, *crtz*, *cruytz*, *crvitz*, *crvtz*, *kraitz*, *kreevcz*, *kreitz*, *kreucz*, *krecz*, *kreuecz*, *kreutz*, *kretz*, *krevcz*, *krevecz*, *krewcz*, *krewtcz*, *krewtz*, *krewz*, *krucz*, *kruecz*, *kruicz*, *kruitz*.

In the entire Reference Corpus of Middle High German[2] (REM, 10501350), there are 83 spelling variants of the word *Teufel* 'devil':

> *dievel*, *diuel*, *diufal*, *diuual*, *diuvil*, *divel*, *divuel*, *divuil*, *divvel*, *dufel*, *duoifel*, *duovel*, *duuel*, *duuil*, *duvel*, *duvil*, *dvofel*, *dvuil*, *dwowel*, *teufel*, *tevfel*, *thufel*, *thuuil*, *tiefal*, *tiefel*, *tiefil*, *tieuel*, *tieuil*, *tieuuel*, *tieuuil*, *tievel*, *tievil*, *tifel*, *tiofel*, *tiuel*, *tiufal*, *tiufel*, *tiufil*, *tiuil*, *tiuofel*, *tiuuel*, *tiuuil*, *tiuval*, *tiuvel*, *tiuvil*, *tivel*, *tivfel*, *tivil*, *tivuel*, *tivuil*, *tivvel*, *tivvil*, *tivwel*, *tiwel*, *tubel*, *tubil*, *tueuel*, *tufel*, *tufil*, *tuifel*, *tuofel*, *tuouil*, *tuovel*, *tuovil*, *tuuel*, *tuuil*, *tuujl*, *tuvel*, *tuvil*, *tvfel*, *tvivel*, *tvivil*, *tvouel*, *tvouil*, *tvovel*, *tvuel*, *tvuil*, *tvvel*, *tvvil*, *tyefel*, *tyeuel*, *tyevel*, *tyfel*

[1] https://www.linguistics.rub.de/anselm/
[2] https://www.linguistics.rub.de/rem/

– minor differences, e.g., in the use of diacritics, are ignored here.

Some of the variance is due to graphemic variation (e.g., *u* vs *v* as in *crutz* vs *crvtz*). Other variants reflect phonetic differences between dialects (e.g., voiced *d* vs voiceless *t* as in *dievel* vs *tievel*).

I provide the full set of variants here to give the reader an impression of the extent and systematicity of the variance. For instance, looking at the variants of *Teufel* 'devil', we see that almost all of the individual word forms follow the general scheme:

1. They all start with a dental consonant (voiced or voiceless: *d*, *t*, *th*),
2. followed by some vowel or diphthong,
3. followed by a labiodental fricative (*u*, *v*, *w*, *f*, or combinations thereof),
4. followed by some vowel,
5. and end with *l*.

The variants of the word *Teufel* that occur in REM cover a surprisingly broad range of the forms that can be generated by the scheme above. For instance, we find *dievel* and *tievel*, but also *divel* and *tivel*, and *dvuil*, *tvuil*, and *diufal*, *tiufal*. But also *tiufal*, *tiufel*, *tiufil* and *tubel*, *tufel*, *tuuel*, *tuvel*, and so on.

In my talk, I want to present some quantitative and qualitative results of spelling variance in historical data of German, but also address variance of morphological and morpho-syntactic features to some extent.

I present two different automatic approaches of normalizing variance, by mapping it either to some artificial form or to modern German. In recent work, we have used the intermediate representations of these approaches – replacement rules and Levenshtein-based mappings – for investigating diatopic variation. First results from these investigations will be presented.