Computing Distance and Relatedness of Medieval Text Variants from German

Stefanie Dipper and Bettina Schrader

Abstract. In this paper, we explore several ways as to computing similarity between medieval text variants from German. In comparing these texts, we apply methods from word and sentence alignment and compute cosine similarity based on character and part-of-speech ngrams. The resulting similarity (or distance) scores are visualized by phylogenetic trees; the methods correctly reproduce the well-known distinction between Middle and Upper German ("Mitteldeutsch" vs. "Oberdeutsch").

1 Introduction[1]

Research on historical languages and their relationships has long been exclusively in the realm of comparative linguistics. For instance, since the early 19th century, Indo-European linguistics has dealt with the reconstruction of language families and language evolution in Europe. Dialects have received similar attention, as they tend to be more conservative than the standard language and often preserve properties of earlier language stages that have been abandoned by the standard language. Hence, historical as well as dialectal data can provide hints on language evolution.

Traditionally, historical linguistic research is mainly based on morphological, phonetic and phonological properties. The relationships between the Indo-European languages e.g. have been established on the base of shared inflectional properties. Furthermore, sound changes (e.g. the first and second Germanic consonant shift) have been used to draw even finer distinctions. Similarly, dialect classification mainly depends on phonetic-phonological features, and rarely takes syntactic properties into account.

Language comparison in this spirit is based on specific language data: for sound-based comparison, lists of parallel words in different languages or language stages are usually used, such as the Swadesh list (Swadesh 1955). Comparisons based on (morpho-) syntactic properties usually use lists of syntactic features whose language-specific values are compared with each another.

More recently people have begun to apply clustering algorithms to such data to compute relationships between dialects (e.g. Nerbonne et al. 1996; Spruit 2006),

1. We thank the anonymous reviewers for their suggestions and many helpful comments.

and to apply phylogenetic clustering algorithms from bioinformatics to derive language family trees in the vain of Indo-European linguistics (e.g. Gray and Atkinson 2003). Phylogenetic algorithms are normally fed with "parallel" DNA sequences of different species and compute (minimal) trees of modification events that relate the species. When applying the algorithms to language data, researchers assume that DNA sequences and language strings behave similar with regard to evolution. Despite commonalities such as diversification or extinction of species and languages, however, there are also differences. For instance, language *contact* is assumed to play an important role in language evolution.

As an alternative to lists of words or features, *corpora* can be used as the base of comparison, and there has been a long-standing tradition of applications in computational linguistics that measure similarities between texts, like authorship attribution (Mosteller and Wallace 1964), and information retrieval and text classification in general. In such an application, a set of features, like word frequency lists or part-of-speech ngrams, is extracted from a (usually large) collection of reference texts, and subsequently used as base of comparison whenever a new text is being classified.

In our paper, we are comparing historical dialect data using such corpus-based methods. However, as historical or dialectal data typically lack (electronically-available) corpora, we are doing without a large collection of reference data. Instead, we base our study on *parallel* corpora, i.e. text variants that share the same underlying "source" text. Hence, quantitative methods can be applied sensibly even to small amounts of data.

In a similar vain, Lüdeling (2006) reports on a study of a (tiny) parallel corpus of Lord's Prayer in five language stages and dialects of German: In this study, Lüdeling (2006) enriches the manually-aligned corpus with phonetic and syntactic annotations and uses these annotations as the basis of comparison.

We also explore several ways as to compute the similarity between medieval German texts. As phonetic annotations along the lines of Lüdeling (2006) require profound knowledge, sound intuitions and a lot of manual work, we decided to compare "writing dialects" ("Schreibdialekte") instead. Just as (a series of) phonetic similarities do not occur by chance, (a series of) spelling similarities can also be used as indications of similarity. In contrast to the study by Lüdeling (2006), we do not use data that has been (manually) pre-aligned. We use an automatic word alignment procedure, and compute similarity based on the alignment results, and character and part-of-speech ngrams. The resulting similarity measures are used to generate phylogenetic similarity trees. Our work presents a pilot study of a much larger project and, hence, is restricted to five tiny texts, each consisting of around 230 words.

The paper is organized as follows. In Sec. 2, we describe the corpus. Sec. 3 and Sec. 4 present the computation of similarity scores based on word alignment and ngram models, respectively. Sec. 5 presents the conclusion.

1.	OMD ("Ostmitteldeutsch"): from Thuringia	MG, 15th
2.	NURN ("Nürnberg"): from the city of Nürnberg (Bavarian, but close to the Franconian border)	UG, 14th
3.	REG ("Regensburg"): from the region of Regensburg (Bavarian)	UG, 15th
4.	AUG ("Augsburg"): from the region of Augsburg (Bavarian, but close to the Swabian border)	UG, 15th
5.	SCHW ("Schwäbisch"): from the region around Kehl (Swabian, but shows alemanic properties)	UG, 15th

Table 1. The five texts of our pilot study and their dialects, along with their larger dialect family (MG: *Middle German*, UG: *Upper German*) and date of origin (century).

2 The Corpus

The texts we used in our pilot study ultimately go back to a Latin source, "Interrogatio Sancti Anselmi de Passione Domini" ('Questions by Saint Anselm about Lord's Passion'). The text consists of a collection of questions posed by Saint Anselm of Canterbury and answered by Saint Mary. In the 14th–16th centuries, this text has been written up in various German dialects (from Upper, Middle, and Low German), and transformed into long and short prose and lyric versions. In total, there are more than 50 manuscripts and prints, which makes the text an exceptionally broadly-documented resource.[2]

The basis of our comparisons are ASCII-transcriptions of the manuscripts, kindly provided to us by the "Altgermanistik" group at Ruhr-Universität Bochum. The transcriptions render the manuscripts as original as possible, so that virtually no interpretation is involved in the transcription process. For our pilot study, we selected five texts from different regions, dialects, and times (see Table 1), and extracted two of Anselm's questions; average length of the extracts is 234.4 words. According to their classifications as *Middle* vs. *Upper German*, one would expect high similarity scores for Texts 2–5 as opposed to Text 1. Texts 2–4 are Bavarian texts, so they should form a cluster as well; however, as noted in Table 1, only Text 3 is a "pure" Bavarian text.

The texts indeed represent parallel texts, e.g., (manually) aligning corresponding words is straightforward, see below the beginnings of Saint Mary's answer to the first question, according to the five texts.[3]

2. A comparable parallel corpus is "The 24 Elders" by Otto of Passau, which is documented by more than 120 manuscripts (not electronically available). Besch (1967) uses 68 of these for a detailed (manual) comparison of their dialects. The comparison focuses on the question which of the dialectal form variants made it into modern German.
3. Rough translation: 'as my dear child had eaten the Last Supper together with his disciples before his martyrdom'. The use of special characters in the ASCII-based transcriptions are explained below.

1. OMD Do mein lieber Son jhe$u$z Da$z nachtmal mit $ienen jüngern am heiligen grün dorn$tage ge$$en hatte
2. NURN Do mein kint mit $einen iungernn het ge e$$en vor $einer marter das iüng$t e$$en
3. REG da mei\- libis chind wol gee$$enn mit $eine\- Junger\- vor $einer marter daz le$t mal
4. AUG Do meín chind híet geezzen. mít $eínen Jungern. vor $eíner marter daz íungí$t mal.
5. SCHW Do min kint hatte gezen daz ivnge$te maz mit sinen ivng'n vor sin' mart'.

The example sentences exhibit interesting peculiarities, e.g. with respect to the vocabulary: in four of the five texts, Mary uses *Kind* 'child' to refer to Jesus, in one text only she uses *Sohn* 'son'; the Last Supper is called *Nachtmahl, jüngstes Essen, letztes/jüngstes Mahl*. The spelling variations *chind/kint* probably reflect phonetic differences; on the other hand, the variations *gee$$en, geezzen* could be solely due to different writing conventions: rendered more closely to the original manuscripts, *$* would look like medial long *ſ*, *z* represents the so-called "tailed z", which looks like ʒ. The combinations *ſſ* vs. *ʒʒ* do not necessarily involve a phonetic difference: *ſſ* is usually interpreted as a writing variant replacing former *ʒʒ*. A similar case is *u/v* alternation, as in *Jungern/ivngern* 'disciples'. Capitalization at this time is used rather arbitrarily: capitals sometimes mark the sentence beginnings or highlight certain prominent nouns (or parts thereof) like *Mutter gottes* 'Mother of God' (not consistently throughout the texts, though).

Variations of the auxiliary 'had' (*hatte, het, híet*) and the participle 'eaten' (*gessen* vs. *ge-essen*) presumably indicate (phonetic-)inflectional differences. With regard to syntax, one can observe that Text 1 (OMD) is most similar to modern standard German: it shows the order V>AUX (as is also usual in subordinate clauses in modern German), and all verb complements precede the head verb. In contrast, NURN, AUG and SCHW show AUX>V (the auxiliary is missing in REG), and they either extrapose all (non-subject) complements to the right (REG, AUG, SCHW) or just one complement appears preverbal (NURN).

These differences can be due to various reasons: they either reflect "truly" linguistic differences (such as the phonetic, inflectional, and syntactic variations), or they can be due to different spelling or writing conventions (*ʒʒ* vs. *ſſ*; capitalization). A certain amount of the text properties, however, is not genuine but "inherited": these properties are due to "translationese" that occurred when one author copied, or even translated, the text from one dialect into the other. In our current study, we have not yet addressed this issue.

Finally, we pre-processed and slightly normalized the texts:

Punctuation: all punctuation marks are deleted (only AUG and SCHW show systematic use of punctuation marks).

Capitalization: all letters are converted to lower case.

Superscripts: superscripts as in $l\overset{i}{v}te$ are rendered as *lv\ite* in the transcriptions. In general, such letter combinations can encode diphthongs or monophthongs (similar to *Umlaut* marking). To ease further processing, we replaced these combinations by the equivalent diphthong letters, e.g. *lvite*.

Abbreviations: all notational abbreviations are spelt out.[4] These include the superscribed horizontal bar ("Nasalstrich") as in *mei\-* (REG), which we replace by *mein*. A superscribed hook ("er-Kürzung") abbreviates *-er*, as in *mart'*, which is replaced by *marter* (SCHW). A frequently-used abbreviation is *vn\-* for the conjunction *und/unde* 'and', which we also eliminated.[5]

3 Similarity Based on Word Alignment

As a means to directly compare the vocabularies and word forms of the five dialects, we aligned them both on the sentence and word level. First, we aligned all ten language pairs (OMD↔NURN, OMD↔REG, etc.) manually on the sentence level. The word alignment has been created automatically using Levenshtein Distance, where deletions, insertions and substitutions of characters are equally punished, and a best-first search computes a full alignment path. Thus, the similarity scores of the aligned word pairs directly reflect their graphemic similarity, and can be analyzed along these lines.[6]

3.1 Quantitative Data Analysis

The common statistic characteristics of the 10 language pairs show that all dialects appear to be highly similar on the graphemic and lexical level (see Table 2): in all

4. Of course, the specific use of abbreviations is also part of a writing dialect. We opted for using full forms to support automatic alignment.
5. In order to know which of the alternative full forms (*und, unde, vnd, vnde*) would have to replace the abbreviated form, we counted all full occurrences in the texts: OMD uses 43x *vnde* and 1x *vnd*, NURN 58x *vnd* and no other variant, etc. Unfortunately, SCHW only uses the abbreviated form, so we had to introduce an artificial "neutral" letter for the full forms in SCHW: *vnd@*.
6. The extracts were too small to use an off-the-shelf statistical alignment tool like GIZA++ (Brown et al. 1990, 1993; Och 2000). Instead, we used the functionality provided by the hybrid alignment tool ATLAS: it has been designed specifically to use a variety of alignment strategies and to align even extremely small parallel corpora (Schrader 2006).

Lang. Pair	Minimum	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
OMD-NURN	0.29	0.67	0.83	0.80	1.00	1.00	0.17
OMD-REG	0.20	0.66	0.78	0.78	1.00	1.00	0.18
OMD-AUG	0.40	0.63	0.73	0.75	0.88	1.00	0.17
OMD-SCHW	0.29	0.67	0.82	0.80	1.00	1.00	0.17
NURN-OMD	0.29	0.67	0.83	0.80	1.00	1.00	0.17
NURN-REG	0.27	0.77	0.90	0.86	1.00	1.00	0.17
NURN-AUG	0.36	0.70	0.85	0.82	1.00	1.00	0.16
NURN-SCHW	0.40	0.75	0.86	0.85	1.00	1.00	0.14
REG-OMD	0.20	0.66	0.78	0.78	1.00	1.00	0.18
REG-NURN	0.27	0.77	0.90	0.86	1.00	1.00	0.17
REG-AUG	0.25	0.71	0.83	0.81	1.00	1.00	0.17
REG-SCHW	0.30	0.71	0.83	0.81	1.00	1.00	0.16
AUG-OMD	0.40	0.63	0.73	0.75	0.88	1.00	0.17
AUG-NURN	0.36	0.70	0.85	0.82	1.00	1.00	0.16
AUG-REG	0.25	0.71	0.83	0.81	1.00	1.00	0.17
AUG-SCHW	0.43	0.67	0.82	0.80	0.94	1.00	0.15
SCHW-OMD	0.29	0.67	0.82	0.80	1.00	1.00	0.17
SCHW-NURN	0.40	0.75	0.86	0.85	1.00	1.00	0.14
SCHW-REG	0.30	0.71	0.83	0.81	1.00	1.00	0.16
SCHW-AUG	0.43	0.67	0.82	0.80	0.94	1.00	0.15

Table 2. Statistics on Word Alignment

cases, the mean of the similarity scores, computed over all word alignments in a language pair, ranges between 0.75 and 0.86. Similarly, the standard deviation is relatively small in all cases. The median, however, shows a more ragged distribution with a minimum value of 0.73 and a maximum of 0.90, which we take as indication that there are systematic differences between the five dialects. Still, when taking the statistics at face value, all five dialects appear to be highly similar.

However, if we rank the language pairs using the mean, we can hypothesize that on average, the OMD text is most dissimilar to all other texts of our case study: in all pairwise comparisons, the pairing involving OMD receives a mean that is considerably lower than the means of all other language pairs. NURN represents the opposite of OMD by receiving the highest means in all pairings, i.e. it appears to be most similar to all other texts. We observe further that the three Bavarian texts NURN, REG, and AUG appear not to be overly similar.

An analysis of the standard deviations basically confirms these findings. We assume that a low standard deviation reflects a high similarity between languages, based on the intuition that if two languages are, on the whole, very similar, then the variation around the mean similarity score should be low, too. Thus, OMD is again clearly shown to be most dissimilar to all other languages of our sample, language

pairs involving OMD always having the highest standard deviations. On the opposite extreme, NURN and SCHW have the lowest standard deviation of all language pairings, thus we may hypothesize that the vocabularies of these dialects, at least in our corpus, are most similar.

3.2 Qualitative Analysis

Following the statistical analysis of the word alignment, we also inspected some of the data qualitatively, in order to (i) evaluate the quality and hence reliability of the automatic word alignment, and (ii) analyze which types of graphemic variation occurred, and hence assess whether our automatic procedure is suitable to give insights into how similar or different the involved dialects are.

With respect to the alignment quality, we exploited part-of-speech information (see Sec. 4) to extract possibly erroneous word pairs: if a word pair shows part-of-speech mismatches, as e.g. in example (1), we have reasons to assume that these category mismatches are either due to real category changes, and, hence, possible divergences between the two languages, or due to erroneous word alignment. In sum, we found out that 12.44% (254) of all word pairs (2042) showed category mismatches. Most of these are indeed due to alignment errors (185, or 9.06%). Others, however, point to minor tagging mismatches as in example (2), which may or may not reflect grammatical differences.

(1) *leut* (noun) – *waren* (auxiliary)

(2) *die* (pronoun) – *dí* (determiner)

Exemplarily, we also analyzed all those word pairs in the language pair OMD–NURN[7] that had a similarity of 0.8 or above (excluding exact matches, i.e. word pairs with a similarity of 1).

In this small data sample, we found 40 word pairs that are overall similar, but simultaneously show systematic variation:

(3) *zu/tzu, juda$z/judas, jo$zeph/io$eph, kinde$z/kindes*

(4) *wil/will, hatten/heten, fur$ten/für$ten, $i/$ie, gieng/ging, bruderen/prudern*

(5) *girig/geitig* (lexical), *sach/gesach* (morphological)

(6) *vnde*, *vnd* and *vnn*; *pfennige* and *phennige*

Probably, the examples in (3) can be interpreted as being due to different writing conventions. On the other hand, the examples in (4) most probably reflect real phonetic

7. Assuming that the dissimilarities of this dialect pair, as statistically shown above, should allow us to gain valuable insights.

differences. We also observed word pairs that indicate lexical and morphological differences between the two dialects (5). Finally, we found instances of inconsistent spelling within the same text, e.g. the OMD text contains all spelling variations shown in (6). Such inconsistencies could indicate ongoing *changes* in a dialect's system, or changes that have occurred rather recently, resulting in spelling uncertainties.

Although the above analysis is rather sketchy, and restricted to very few word pairs, they are sufficient to highlight the usefulness of our methodology: without investing too much effort in a sophisticated procedure, we are able to automatically word-align small text samples of historical dialects. We are using a very simple similarity measure that relies exclusively on the graphemic similarity of the tokens in the parallel corpus. The error rate of the word alignment procedure seems to be quite low. Furthermore, the similarity measure also provides means for (i) comparing the dialects as a whole, in order to determine whether on principle, two dialects are similar, and (ii) to extract remarkable word pairs from the corpus for a detailed linguistic analysis.

4 Similarity Based on Ngrams

In addition to computing similarities based on word-alignment measures, we ran another series of experiments based on frequencies of character ngrams (4.1) and part-of-speech tag ngrams (4.2). The ngram models serve us as *text fingerprints* or *profiles*, which are compared one to another.

4.1 Character Ngrams

Ngram statistics are widely used in natural language processing. Most often, ngrams consist of sequences of words or part-of-speech tags, less frequently ngrams of *characters* are used. Character-based ngram models have been applied to a range of tasks that involve similarity computations, such as spelling error detection (Zamora et al. 1981), authorship attribution (Fuchun et al. 2003; Kjell 1994), language or topic classification (Beesley 1988; Cavnar and Trenkle 1994), and information retrieval in general (Damashek 1995).

In our study, we compared the distributions of character ngrams, with n=1,2,3, across the texts. As an example, Table 3 lists the top-ranked character bigrams across the texts, along with their absolute frequencies and their weights, which determine the ranking in the table. $W_{t,n}$, the weight of an ngram n in a text t, is computed as $W_{t,n} = \frac{F_{t,n}}{F_{T,n}}$, where $F_{t,n}$ is the frequency of an ngram in a specific text, and $F_{T,n}$ is the frequency of the ngram in the entire collection (i.e., in all five texts). $W_{t,n} = 1$ means that the ngram only occurs in the current text and, hence, can be interpreted as a

OMD			NURN			REG			AUG			SCHW		
a$	10	1.00	iu	5	0.62	sy	7	1.00	lí	6	1.00	vd	7	1.00
$z	26	1.00	pf	8	0.57	aw	5	0.62	eí	16	1.00	iv	9	1.00
th	10	0.67	as	11	0.50	ai	5	0.62	zz	5	1.00	$v	8	1.00
tt	8	0.67	fe	8	0.50	un	10	0.62	dí	12	1.00	d@	11	1.00
eb	10	0.53	$$	8	0.44	ff	5	0.42	aí	5	1.00	vi	18	0.82

Table 3. Top-ranked character bigrams across the five texts, with absolute frequencies and weights

characteristic feature of this text (and, ideally, this can be related to features specific to the dialect used in this text). That is, with regard to writing conventions, Text AUG and SCHW stand out due to their idiosyncratic spellings. In contrast, Text NURN exhibits a sort of "average spelling" (similar observations have been made in Sec. 3).[8]

4.2 POS Unigrams and Bigrams

Character ngrams can give hints as to whether two texts share a certain amount of vocabulary and (graphemic-)phonetic and inflectional features. They certainly cannot be used for syntactic comparisons. Lüdeling (2006) used syntax trees in her study; in contrast, we use POS ngrams as a cheap surrogate for syntax. We manually annotated all texts with POS tags according to the STTS tagset.[9]

8. The tendencies that one can observe with bigrams also show up with the other ngram types. This is partly related to the way we compute the scores: one of the reasons of, e.g., *d@* being a sequence unique to Text SCHW is the fact that we introduced the letter @ only in this text, see Fn. 5. Similarly, the character *í* only occurs in Text AUG, but with high frequency. A more sophisticated ngram weighting measure would therefore take the frequencies of (n-1)grams into account.

9. `http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.ps.gz`.
The STTS tagset has been developed for the annotation of modern German. In applying the tagset to our texts, we encountered surprisingly few problematic cases: Distinguishing adverbs (ADV) from verb particles (PTKVZ) can be difficult, as in *vnde $ie von dem ti$che **vf** waren ge$tanden* ('and they had stood up from the table', OMD). Pronominal adverbs such as *davon* 'thereof' are mostly spelt in two words, which we annotated as *do/ADV von/APPO*. For seemingly noun compound constructions, we defined a new tag PTKNZ ("Partikel Nomenzusatz", 'noun particle') for the non-head components: *erb/PTKNZ $chafft/NN* 'heir_ship', *nach/PTKNZ kommen/NN* 'off_spring', *an/PTKNZ vanch/NN* 'begin'.
A final example is the distinction between demonstrative and relative pronouns, which in modern German is done on the basis of the verb position (verb second vs. verb final). However, the evolution of the modern verb position is a research question in itself, so no prior decision should be made during POS annotation. A relevant example from our texts is *do kaufften in einer hand leut **die** hie$$en y$mahelite* ('Then some people bought him, who were called Ysmaelite/They were called Ysmaelite.', NURN). We currently use underspecified tags for these cases: *die/PDS_PRELS*.

POS Tag	OMD		NURN		REG		AUG		SCHW	
NN	42	18.03	44	17.96	42	18.10	43	19.20	45	18.91
PPER	22	9.44	26	10.61	26	11.21	21	9.38	25	10.50
ART	23	9.87	19	7.76	20	8.62	20	8.93	20	8.40
ADV	16	6.87	20	8.16	15	6.47	16	7.14	21	8.82
VVFIN	13	5.58	18	7.35	17	7.33	16	7.14	17	7.14
VAFIN	16	6.87	16	6.53	14	6.03	14	6.25	15	6.30
Sum	233	100	245	100	232	100	224	100	238	100
ART+NN	21	8.97	15	6.10	17	7.30	17	7.56	17	7.11
PPOSAT+NN	3	1.28	10	4.07	7	3.00	9	4.00	9	3.77
ADJA+NN	8	3.42	2	0.81	6	2.58	4	1.78	4	1.67
Sum	234	100	246	100	233	100	225	100	239	100

Table 4. Top-ranked POS unigrams (top) and bigrams (bottom) across the texts, along with their absolute and relative (%) frequencies

One way of exploiting the POS information would be to align these tags across the texts and compute similarities just as described in Sec 3. Another way, however, is to explore properties of the POS annotations themselves and compare *patterns* of POS annotations across the texts. In a similar vein, Lauttamus et al. (To Appear) applied this method to the classification of language data produced by first vs. second language learner.

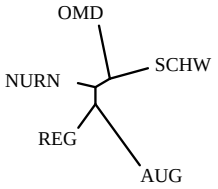
POS ngrams distribute much more evenly across the texts than character ngrams, so we can present an overview of the rankings across all texts. Table 4 lists the top-ranked POS unigrams and bigrams, along with their frequencies.

As can be seen from Table 4, all texts except OMD show similar rankings with regard to unigrams and bigrams: ranks 1-2 are occupied by NN and PPER, respectively. Ranks 3–5 are taken by ART, ADV, and VVFIN, in slightly varying orders. With bigrams, Texts 2–5 show identical rankings. In contrast, OMD ranks ART second rather than PPER, and prefers ADJA+NN over PPOSAT+NN among the bigrams (these tag sequences occur e.g. in the formulae *mein/PPOSAT kint/NN* ('my child') and *mein lieber/ADJA Son/NN* ('my dear son').

4.3 Computing Similarity and Relatedness

The character and POS ngrams are used to compute pairwise similarity between the texts. For computing similarity, we applied the cosine measure. Pairwise similarity measures result in a similarity matrix, cf. Table 5. Based on the similarity matrix, we computed a phylogenetic tree, using the *Neighbor Joining Method* (Saitou and Nei 1987). This algorithm first relates the least distant pair, merges them, and applies the

Char. Bigrams	OMD	NURN	REG	AUG	SCHW
OMD	100	88.76	85.93	80.79	88.61
NUERN		100	93.44	85.31	91.40
REG			100	87.27	86.12
AUG				100	82.47
SCHW					100



POS Bigrams	OMD	NURN	REG	AUG	SCHW
OMD	100	80.73	85.79	84.69	83.67
NUERN		100	92.80	93.18	92.84
REG			100	95.56	93.94
AUG				100	93.25
SCHW					100

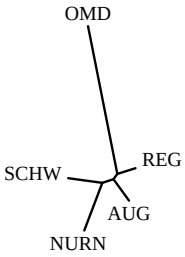


Table 5. Similarity matrices for pairwise similarities based on character (top) and POS (bottom) bigrams and the corresponding phylogenetic (unrooted) trees

algorithm to the remaining pairs plus the newly created merged one. The result can be visualized, e.g., by unrooted trees. Table 5 shows the similarity matrices along with the corresponding phylogenetic trees that cluster character and POS bigrams.[10]

The results shown in Table 5 confirm our observations made with regard to the rankings of the top-ranked ngrams in Tables 3 and 4: (i) At the level of character bigrams, AUG and OMD are the most idiosyncratic texts, with values ranging between 80.79–87.27 (AUG) and 80.79–88.76 (OMD). Hence, AUG and OMD are located on rather long isolated branches of the phylogenetic tree, while NURN is the most "average" one (with values of 85.31–93.44), and therefore located in the center of the tree. (ii) At the level of POS bigrams, OMD clearly stands out, while the other texts cluster nicely. With neither of the measures would the three Bavarian texts, NURN, REG, AUG, form a cluster on their own, again as already observed in the alignment experiment, Sec. 3.

10. The length of the branches indicates the degree of distance. The scale used in both figures are different, though. The trees have been created by the software package PHYLIP (Felsenstein 1989).

5 Conclusion

The goal of this paper was to investigate whether quantitative methods can be sensibly applied to small text samples of historic German dialects. We showed first that a simple, automatic word alignment procedure can be used to align the data successfully. We then used the word alignment and character and POS ngram models to detect similarities and differences between these writing dialects, and to compute clusters of dialects. One of the dialects, OMD, was clearly shown to be most dissimilar to the other four dialects. This result correctly reproduces the well-known distinction between Middle and Upper German.

The four Upper German dialects, on the whole, could be shown to be highly similar by the similarity computations based on word-alignment and POS ngrams. In contrast, character ngrams singled out AUG as highly idiosyncratic, closely followed by OMD. The three Bavarian texts could not be shown to form a cluster, which might be attributed to the fact that only REG represents a "canonical" Bavarian text.

Finally, with the string-based methods (word-alignment and character ngrams) NURN came out as the most "neutral" dialect, sharing many characters and character sequences with the other dialects.

Our next steps, in the context of a larger project, will be to expand our data to include up to 50 complete corpus samples. We also plan to repeat our word alignment analysis using a more refined similarity measure which incorporates linguistic knowledge on sound changes ("Lautgesetze"). Similarly, in a sort of bootstrapping approach, we would integrate prior findings (like the correspondance of *$$*–*zz* in certain text pairs) into the alignment procedure and use this information in subsequent alignments of the respective texts.

Finally, it is planned to more thoroughly investigate into the historical origins of the data, especially with respect to authorship and translational routes from one dialectal text to the next. Thus, we hope to determine which similarities of the data are possibly due to "translationese" that occurred when one author copied, or even translated, the texts from one dialect into the other. Ultimately we would like to be able to automatically determine which of the similarities between two texts are artifacts of the copy procedure, and which ones derive from genuine linguistic proximity of the dialects involved.

References

Beesley, Kenneth R. (1988). Language identifier: A computer program for automatic natural-language identification on on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, 47–54.

Besch, Werner (1967). *Sprachlandschaften und Sprachausgleich im 15. Jahrhundert. Studien zur Erforschung der spätmittelhochdeutschen Schreibdialekte und zur Entstehung der neuhochdeutschen Schriftsprache*. München.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin (1990). A statistical approach to machine translation. *Computational Lingusitics* 16(2):79–85.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993). The mathematics of machine translation: Parameter estimation. *Computational Linguistics* 19(2):263–311.

Cavnar, William B. and John M. Trenkle (1994). N-gram-based text categorization. In *Proceedings SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*.

Damashek, Marc (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science* .

Felsenstein, Joseph (1989). PHYLIP — Phylogeny Inference Package (version 3.2). *Cladistics* 5:164–166.

Fuchun, Peng, Dale Schuurmans, Vlado Keselj, and Shaojun Wang (2003). Language independent authorship attribution using character level language models. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary.

Gray, Russell D. and Quentin D. Atkinson (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426:435–439.

Kjell, Bradley (1994). Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing* 9(2):119–124.

Lauttamus, Timo, John Nerbonne, and Wybo Wiersma (To Appear). Detecting syntactic substratum effects automatically in interlanguage corpora. In Muriel Norde, Bob de Jonge, and Cornelius Hasselblatt (eds.), *Language Contact in Times of Globalization*, Amsterdam: Benjamins.

Lüdeling, Anke (2006). Using corpora in the classification of language relationships. *Zeitschrift für Anglistik und Amerikanistik. Special Issue on 'The Scope and Limits of Corpus Linguistics'* 217–227.

Mosteller, Fred and David Wallace (1964). *Inference and Disputed Authorship: The Federalist Papers*. Massachusetts:Addison-Wesley.

Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten, and Willem van de Vis (1996). Phonetic distance between dutch dialects. In *Proceedings of the Sixth CLIN Meeting*, 185–202, Antwerp.

Och, Franz Josef (2000). Giza++: Training of statistical translation models. `http://www-i6.informatik.rwth-aachen.de/~och/software/GIZA++.html`.

Saitou, Naruya and Masatoshi Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406–425.

Schrader, Bettina (2006). Atlas – a new text alignment architecture. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 715–722, Sydney, Australia: Association for Computational Linguistics.

Spruit, Marco Rene (2006). Discovery of association rules between syntactic variables — data mining the syntactic atlas of the Dutch dialects. In *Computational Linguistics in the Netherlands*.

Swadesh, Morris (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21:121–137.

Zamora, E. M., Joseph J. Pollock, and Antonio Zamora (1981). The use of trigram analysis for spelling error detection. *Information Processing & Management* 17(6):305–316.