Bachelor's Thesis

Clustering Dialect Varieties Based on Historical Sound Correspondences

Author
Verena Blaschke
verena.blaschke@student.uni-tuebingen.de

Supervisor
Dr. Çağrı Çöltekin
ccoltekin@sfs.uni-tuebingen.de

A thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Arts
in
International Studies in Computational Linguistics

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

August 2018

Abstract

While information on historical sound shifts plays an important role for examining the relationships between related language varieties, it has rarely been used for computational dialectology. This thesis explores the performance of two algorithms for clustering language varieties based on sound correspondences between Proto-Germanic and modern continental West Germanic dialects. Our experiments suggest that the results of agglomerative clustering match common dialect groupings more closely than the results of (divisive) bipartite spectral graph co-clustering. We also observe that adding phonetic context information to the sound correspondences yields clusters that are more frequently associated with representative and distinctive sound correspondences.

Contents

List of Tables

List of Figures

1 Introduction

Language variation and change have long been a focus of linguistics. The analyses necessary for determining systematic similarities and differences between language varieties had to originally be exclusively performed by hand, but with the advances of computational methods, it has become possible to carry out quantitative analyses more easily. One field concerned with such analyses is *dialectometry*, which focuses on computational and statistical methods for dialectology.

Applying quantitative methods to dialectology gives the advantage that statistical models can work using all the feature information of the data that they are given, quickly evaluating for each feature how well it does or does not describe similarities or differences in the data.

In this thesis, we examine a set of West Germanic language varieties currently spoken in continental Europe. We compare them by investigating how they have changed phonologically since a shared ancestral stage of Germanic from around 2500 years ago. Our goal is to automatically assign a cluster structure to the modern language varieties that reflects shared sound changes within each cluster and differences between sound shifts between different clusters.

This thesis is structured as follows: In the following subsection, we present the dialectometrical approaches that influenced our work. Then we begin by introducing the data in section 2. Next, in section 3, we give a brief introduction to continental West Germanic languages and dialects, as well as proposed ways of sorting them into groups and the problems associated with doing this. In section 4, we explain our methodology for aligning the data and extracting sound correspondences, describe two approaches to clustering the data, and then explain how we rank the sound correspondences associated with each cluster. We present the results in section 5 and discuss them in section 6, before concluding the thesis in section 7.

1.1 Related Work

In the past decades, there have been many advances in the field of dialectometry.[1] The following works are especially relevant in the context of this thesis.

Prokić et al. (2012) perform hierarchical clustering on Bulgarian dialects based on phonetic distances. This is similar to the work by Prokić (2007), wherein she performs an aggregate analysis of the data via an unspecified clustering method based on a dialect-by-dialect matrix storing phonetic distance values, which she compares to individual analyses of recurring sound correspondences between the dialects. The latter analyses are in turn related to the work by Prokić and Cysouw (2013) who explore more closely how to automatically judge the regularity of sound correspondences for investigating dialect transitions in the geographical spread.

[1] For thorough overviews, see Nerbonne (2009) and Wieling and Nerbonne (2015).

Heggarty et al. (2010) worked with modern varieties of Germanic languages. They applied the NeighborNet method (Bryant and Moulton, 2004) based on pronunciation differences to represent the data as a web-like phylogenetic network.

Pröll (2013) also investigated a clustering method that does not use strict hierarchies or categories by applying fuzzy clustering on the basis of lexical variation to capture gradual changes between dialect groups.

Wieling and Nerbonne (2009; 2011) examined the relation between dialect groups and the phonetic properties that categorize them. They extracted sound correspondences between dialects and a reference dialect and used bipartite spectral graph co-clustering for simultaneously clustering sound correspondences and dialects. This method was originally introduced for data mining (Dhillon, 2001; Zha et al., 2001), but it has also been used in bioinformatics (Kluger, 2003). Wieling and Nerbonne (2009; 2011) applied this method to dialects spoken in the Netherlands.

A hierarchical version of this co-clustering method was used by Wieling and Nerbonne (2010), again for dialects spoken in the Netherlands, and Wieling et al. (2013) employed this method for clustering British English dialects. Montemagni et al. (2013) applied this method to Tuscan dialects, and supplemented the sound correspondences with information on the phonetic contexts of the sound segments in the correspondences.

Our usage of phonetic context information is also influenced by the work of Wettig et al. (2012) who used context-sensitive sound correspondence rules for aligning phonetically transcribed data from related languages.

2 Data

We work with phonetically transcribed data from continental European West Germanic (henceforth: CWG) dialects and standard languages (hereafter collectively referred to as doculects). The data we work with are taken from the Sound Comparisons project, an extension of the Languages and Origins in Europe project (Renfrew and Heggarty, 2009), lead by Heggarty (2018), who compiled IPA transcriptions of word lists in a range of Germanic doculects.

From this database, we used 110 cognate sets (also referred to as *concepts*) from 20 modern CWG doculects and a reconstructed version of Proto-Germanic.[2] Of the modern doculects, two are identified as standard languages in the database (Dutch spoken in the Netherlands and Belgium[3]), the rest as local vernaculars. The modern doculects are from locations in the Netherlands, Belgium, Luxembourg, (along the Western border of) Germany, France (Alsace), Switzerland, Liechtenstein, Austria (Voralberg), and Italy (South Tyrol). Figure 1 provides an overview of these locations. The legend is explained in section 3.

[2] The Sound Comparisons project does not state the theoretical basis for the Proto-Germanic reconstruction. According to the project website, the reconstruction might be close to a variant of the language spoken in around 500 BCE in Southern Scandinavia.

[3] Hereafter referred to as *Std. Dutch (NL)* and *Std. Dutch (BE)*, respectively.

For the phonetic alignment step (see section 4.1), we used 14 additional doculects that are Germanic but not CWG. To control for transciber bias, i.e. different transcribers providing slightly different transcriptions of identical sounds, we only worked with doculects that share the same transcriptor, Warren Maguire. The transcriptions of the modern doculect data are narrow transcriptions; that of the Proto-Germanic reconstruction appears to be broader.[4]

The concepts are often represented in root forms of the words to mitigate the overrepresentation of certain affixes (Renfrew and Heggarty, 2009).[5]

We excluded one CWG doculect that covered only 35 concepts. The Proto-Germanic data cover all 110 concepts; each of the modern doculects covers at least 103 concepts, and each concept is covered by at least 17 modern doculects. In total, we have 2181 word alignments between Proto-Germanic and modern CWG doculects.

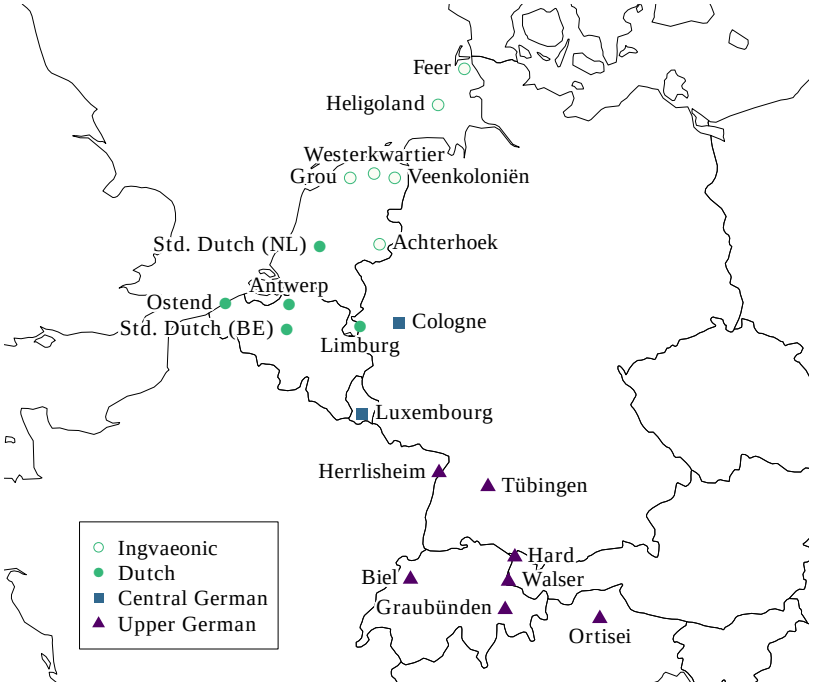

Figure 1: Locations of the modern continental West Germanic doculects we worked with.

[4] For instance, the Proto-Germanic data include no suprasegmentals.
[5] E.g., verbs are represented in their imperative forms rather than the infinitive.

3 Continental West Germanic

The CWG doculects include several standard languages (standard varieties of Dutch spoken in Belgium and the Netherlands, Luxembourgish, standard varieties of German in Germany, Austria, Switzerland and Liechtenstein) as well as many regiolects and dialects. Establishing subgroups within this collection of doculects provides a challenge that has been taken up many times, with different results. Even the classification of West Germanic as its own branch of Germanic is controversial, though generally accepted (Voyles (1971); Harbert (2007, pp. 7-8); Ringe (2012)).

Within the CWG group, it gets even more complicated and contested. Nielsen (1989, pp. 72-80) gives an overview of the history of attempts to divide the West Germanic dialects into subgroups with the associated criteria (phonological, morphological, lexical, and/or extra-linguistic) and criticisms.

Much of the challenge of grouping CWG doculects stems from them being very similar to one another and closely related. These similarities do not only exist because of genetic relatedness but also—enabled by the geographic proximity—mutual influences (Harbert, 2007, p. 8).

On the other hand, interactions between dialects and standard languages have also influenced the dialect landscape (van Coetsem, 1992). Kremer and Niebaum (1990) found that in Germany and in the Netherlands (but not in Switzerland), dialects tend to become closer to the standard languages, with the result of state or standard language borders tending to act as dialect borders.[6]

Heggarty et al. (2010) describe models for intra-family variation, most importantly two major models: the tree-like, hierarchical *splits* model, and the *waves* model, which corresponds more closely to a dialect continuum.

A combination of the two is reflected in Figure 2, which shows a proposed division of CWG doculects into three main groups: North Sea Germanic (including Frisian and Low German), Franconian (including Dutch and High Franconian) and Alpine Germanic (including Alemannic and Bavarian), and presents High German as the result of the convergence of High Franconian, Alemannic and Bavarian.

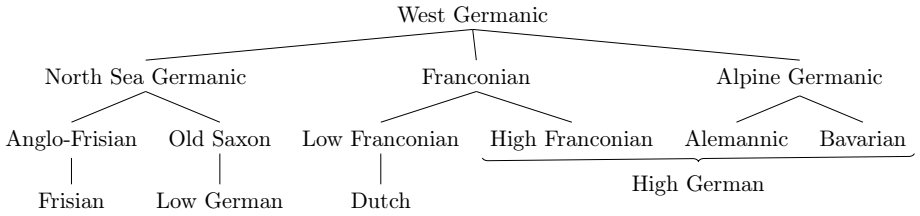

Figure 2: The internal structure of continental West Germanic based on Harbert (2007, p. 8).

[6] They also found that Low German dialects and CWG spoken in Non-Germanic regions tend to be replaced by the prevailing standard language instead.

Heggarty et al. (2010), who have inspected the same data that we work with, describe their results as "a progressive dialect continuum [...] incrementally proceeding in fairly close step with geography."

Alternatively, Hammarström et al. (2018), whose language catalogue Glottolog contains strictly hierarchical categorizations, give an entirely tree-like classification of the CWG doculects, as shown in Figure 3. This classification is based on the work by Stiles (2013) and, like the previous figure, Harbert (2007). We include it here since the output of our clustering methods is also strictly hierarchical.

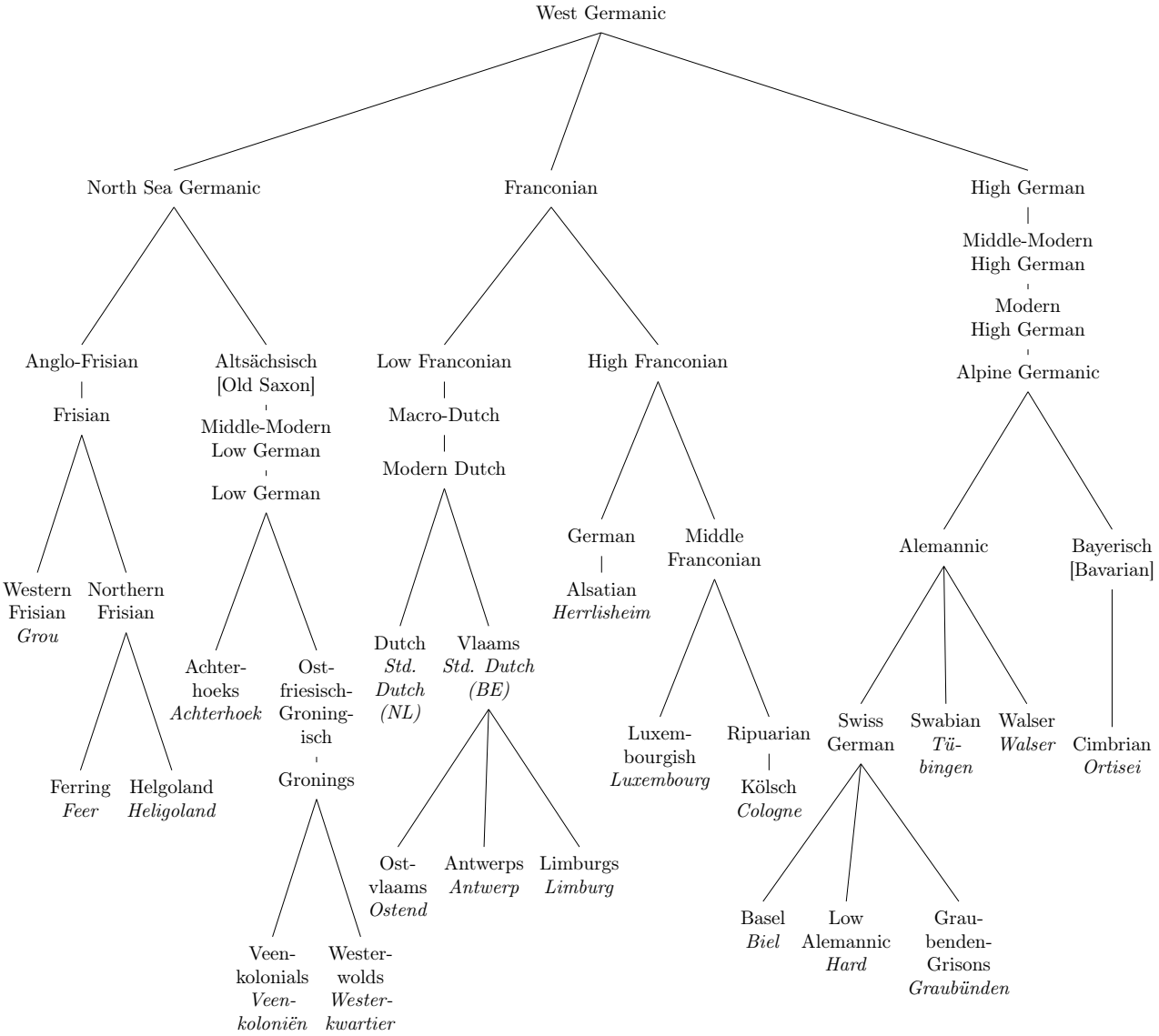

Figure 3: The full classification tree (up to West Germanic) for the modern doculects we used as defined by Hammarström et al. (2018). The names of the modern doculects are dispayed in italics.

3.1 North Sea Germanic

Stiles (2013) posits that the most significant division of West Germanic varieties is the split into Ingvæonic (that is, North Sea Germanic) varieties and non-Ingvæonic varieties. This split is also supported by, e.g., Harbert (2007, p. 7), Sonderegger (1979, pp. 117–123) and van der Auwera and Van Olmen (2017). What is more complicated is defining which doculects are Ingvæonic:

> Stiles (2013) defines this group as Frisian, English, and "to a certain extend, Old Saxon" (i.e., Low German).
>
> Harbert (2007, pp. 7–8, 17) defines it as Frisian, English, and Low German, while noting that Dutch has also been influenced by the Ingvæonic languages.
>
> Sonderegger (1979, pp. 71, 117–123) classifies Ingvæonic as Frisian, English, Low German and (having become a part of this group more recently) Dutch.
>
> Van der Auwera and Van Olmen (2017) define Ingvæonic as Frisian, English, and Dutch.

The distinct properties of the Ingvæonic subgroup concern mostly inflection and pronouns (Stiles (2013); Harbert (2007, pp. 7-8)), although Stiles (2013) also lists some phonological characteristics: "backing of long and short $*a$ before nasals [...]; fronting of long and short $*a$; and palatalization of velar consonants".

We follow the categorization by Harbert (2007), as it reconciles most of the aforementioned classification options. A split similar to the one proposed by Sonderegger (1979) is part of the following section. We thus divide the modern doculects from our dataset as follows:

> Ingvæonic: Feer, Heligoland, Grou, Westerkwartier, Veenkoloniën, Achterhoek
>
> Non-Ingvæonic: Std. Dutch (NL), Std. Dutch (BE), Ostend, Antwerp, Limburg, Herrlisheim, Luxembourg, Cologne, Ortisei, Tübingen, Walser, Biel, Hard, Graubünden.

In Figure 1, the Ingvæonic doculects are marked with green-rimmed, non-solid circles.

3.2 Results of the High German Sound Shift

A very important development for some of the CWG doculects, especially High German, is the High German sound shift.[7] Summarizing Harbert (2007, pp. 47–48) and König (2005, pp. 62–64), we can outline the High German sound shift as follows:

[7] The term *High German sound shift* has been used both to describe the High German consonant shift, and to describe the consonant shift as well as sound shifts concerning the High German vowel system. We use it here with the former meaning.

The voiceless (aspirated)[8] Germanic stops (/*p, *t, *k/) underwent lenition and shifted into affricates or fricatives. (Typically, these stops developed into fricatives in postvocal positions, and into affricates in word-inital or postconsonantal positions. Moreover, /*t/ changed more commonly than /*p/, which in turn changed more commonly than /*k/.) To balance this out, the voiced Germanic stops (/*b, *d, *g/) on the other hand developed into their voiceless (and aspirated) counterparts.

Generally, these changes are more pronounced in the Southern CWG area, and did not take place in the North (Noble, 1983, p. 33)[9]. In between, there are many doculects that only partially realized the High German sound shift, with some of the changes only applying to individual words (König, 2005, p. 63).

Based on this, there is a common division of CWG doculects spoken in Germany into three groups: Upper German doculects, which almost completely exhibit lenition for all three voiceless stops (except for sometimes /*k/ > /(k)x/), Central German doculects, which show a partial development of the High German sound shift, and Low German doculects, which were not influenced by the High German sound shift (Noble, 1983, pp. 33, 55).

This division is performed based on the presence or absence of this shift in individual words (König, 2005, p. 63). The pronunciation boundaries (*isoglosses*) for such words sometimes appear tightly bundled together, although such bundles can also fan out such that a region contains a continuum of very subtle dialect differences, as is the case for the so-called Rhenish Fan at the Western part of the boundary (or transition zone) between Low and Central German (König, 2005, pp. 63, 138, 141).

We base the following classification of the CWG doculects we worked with on maps by König (2005, pp. 64, 230–231):[10]

> Low German, Dutch, and Frisian: Westerkwartier, Veenkoloniën, Achterhoek, Feer, Heligoland, Grou, Std. Dutch (NL), Std. Dutch (BE), Ostend, Antwerp and Limburg.
>
> Central German: Cologne and Luxembourg.
>
> Upper German: Tübingen, Herrlisheim, Biel, Graubünden, Walser, Hard and Ortisei.

Figure 1 shows this division: Low German, Dutch and Frisian are marked with circles, Central German with blue squares and Upper German with purple triangles.

This division also matches the intra-database grouping by Heggarty (2018) (who

[8] Aspiration is not marked in the reconstructed version of Proto-Germanic we worked with.

[9] It is therefore generally assumed that the locations from which these changes spread are in the Southern CWG area, although there are also some controversies surrounding this (Goblirsch, 2005, pp. 155–181).

[10] Central German is delimited to the North (Low German) with an isogloss bundle containing words exhibiting the absence (Low German) or presence (Central German) of affrication or spirantization for /p/ (e.g. *schlapen / schlafen* 'sleep'), /t/ (e.g. *Tid / Zeit* 'time', *Water / Wasser* 'water') and /k/ (e.g. *maken / machen* 'make'). The isogloss bundle serving as boundary between Central and Upper German focuses on the a□rication of /p/ in Upper German (e.g. *Appel / Apfel* 'apple').

additionally split up the first group into Low German on the one hand, and Frisian, Dutch and Flemish on the other).

4 Methods

In section 4.1, we describe how we align the phonetic transcriptions from our data. From the aligned data, we extract sound correspondences (section 4.2), which we then use for two different clustering methods (section 4.3).

We implemented these methods in Python making use of several libraries for statistical analyses: NumPy (Oliphant, 2006), SciPy (Jones et al., 2001), scikit-learn (Pedregosa et al., 2011) and LingPy (List et al., 2018).

4.1 Multiple Sequence Alignment

We carry out alignment based on data from all the investigated doculects at once using multiple sequence alignment. Doing this instead of performing pairwise alignment between the Proto-Germanic and the modern data makes it possible—in addition to using patterns found in doculect-specific sound correspondences—to base the alignment on commonalities between the modern doculects. Because of this, we use all of the modern Germanic data we extracted from the Sound Comparisons project instead of only the CWG doculects.

We use a library-based version (Notredame et al., 2000) of the progressive multiple sequence alignment method (Thompson et al., 1994). For each concept:

1. We divide the phonetic representation of each word into an array of sound segments. These sound segments are typically single IPA tokens (plus diacritics), but we use multi-token segments for affricates, diph- and triphthongs and geminates.[11]
2. We then generate alignments for all possible pairwise combinations of (modern or historical) doculects. These alignments are created using the algorithm by Needleman and Wunsch (1970), with a scoring scheme based on the sound classes introduced by List (2012).[12] All segment alignments from this step are stored in a so-called *library*, each associated with a weight reflecting its relative frequency.
3. We create a sequence-by-sequence distance matrix from the similarity information between each pair of aligned sequences that was used by the scoring scheme in the previous step. We convert the distance matrix into a tree using the UPGMA method (Sokal and Michener, 1958).[13]

[11] Allowing multi-token segments differs from the method employed by, e.g., Wieling and Nerbonne (2010). They neither allowed multi-token segments nor did they add contextual information (see section 4.2), but they remark on a common alignment ∅:[ʃ], which frequently appears after [t]:[t]. We opt instead to interpret affricates as single segments with the result of correspondences such as [t]:[t͡ʃ].

[12] The sound classes are elaborated upon in section 4.2.

[13] This method is explained in section 4.3.1.

4. Progressing from the tips of the tree to the root, we consecutively join the alignments meeting at branchings based on the library created in the first step, until (at the root) all alignments have been consolidated into one alignment table.

We use the LingPy library for Python (List et al., 2018) to perform these steps. Table 1 shows an excerpt from the multiple sequence alignment for the concept "cold".

Doculect	Sound segments					
Proto-Germanic	k	a	l	d	a	z
Westerkwartier	kʰ	o	ë	tʰ	-	-
Luxembourg	kʰ	aː	l	-	-	-
Biel	X	AU	-	t	-	-
Walser	x	aː	l	t͡s	-	-
Ortisei	kʰ	0	l	t͡s	-	-

Table 1: An excerpt from the aligned sequence table for the concept "cold".

4.2 Sound Correspondence Extraction

After performing sound segment-wise alignment, we extract sound correspondences between Proto-Germanic and each modern doculect from the alignment tables for all concepts. We use straightforward segment-to-segment correspondences as well as correspondences that include contextual information:

No context: These are simple segment-to-segment correspondences.

Simple context: We (separately) add information about the left and right single-segment context, stating whether the context is a consonant or a vowel. This can only be performed when the context in question is of the same type for both Proto-Germanic and the modern doculect.

Sound class-based context: This is similar to the previous category, but we give more fine-grained information about consonants and vowels. We use the sound classes introduced by List (2012), which discern between fifteen consonant groups and six vowel groups.

Word boundaries: When the (left or right) context is a word boundary, we add information about this.

Table 2 provides an overview of the different context types, with the corresponding IPA characters found in our data in the case of List's sound classes. IPA characters with diacritics are classified like their diacritic-less counterparts, and diphthongs and triphthongs are classified according to the first character in the sequence. Table 3 shows the sound correspondences that can be inferred for the aligned segments from Proto-Germanic and Ortisei German for the alignment shown in Table 1.

Context type	Abbr.	Definition	IPA characters
Simple context	cons	consonants	
	vow	vowels	
Sound class-based context	A	unrounded open vowels	a, A
	B	labial/labiodental fricatives	f, pf, v, F, B
	C	dental/alveolar affricates	dz, ts, tS
	D	dental fricatives	ð, T
	E	unrounded mid vowels	e, æ, 5, @, E, 3, 2
	G	velar/uvual fricatives	x, X, G
	H	laryngeals	h, H, P
	I	unrounded close vowels	i, ɪ
	J	palatal approximants	j
	K	velar/uvular plosives/affricates	k, kx, q, ¥
	L	lateral approximants	l, ë, ï
	M	labial nasals	m, M
	N	(non-labial) nasals	n, ŋ, M, ð
	O	rounded open vowels	6
	P	labial plosives	b, p
	R	trills/taps/flaps	r, ô, R, ö, K
	S	sibilant fricatives	s, z, ç, S, Z, J
	T	dental/alveolar plosives	t, d, ú
	U	rounded mid vowels	o, ø, œ, 0, 8, ×
	W	labial approximants/fricatives	w
	Y	rounded close vowels	u, y, U, Y
Word boundaries	#	word boundaries	

Table 2: Context representations by context type. For the sound class-based context information (List, 2012), the corresponding IPA characters appearing in our data are included.

Pr.-G.	Ort.	No context	Simple context	Sound class-based context	Word boundaries
k	k^h	k > k^h	k > k^h / _vow		k > k^h / #_
a	0	a > 0	a > 0/ cons_ a > 0 / _con	a > 0/ K_ a > 0 / _L	
l	l	l > l	l > l / vow_ l > l / _cons	l > l / A_	
d	ts	d > ts	d > ts / cons_	d > ts / L_	
a		a > ;	a > ; / cons_		
z		z > ;			z > ; / _#

Table 3: Proto-Germanic–Ortisei German sound correspondences extracted from the aligned entries for the concept "cold".

The context information we use is different from the approach by Montemagni et al. (2013) in that they only distinguished between consonants, vowels, glides, gaps and word boundaries. Moreover, they included left and right context information simultaneously. They also added context information when it is different for the reference doculect and the doculect used for clustering, whereas we present context information in a style more similar to the phonological rewrite rules introduced by Chomsky and Halle (1968).

We ignore gap-gap alignments, as they do not contain information on correspondences between Proto-Germanic and the modern doculect in question, only about inserted sound segments in one or more other doculects. Furthermore, we treat insertions and deletions that LingPy flags as swaps (metathesis) as normal insertions or deletions, as such cases only happen for 3 of the 110 concepts.

For each doculect, we ignore sound correspondences that occur fewer than three times across all concepts to reduce the effect misalignments might have.

After extracting the sound correspondences for all modern doculects, we have a doculect-by-correspondence matrix storing the absolute frequencies of the sound correspondences.

4.3 Clustering

We implemented two approaches to custering the data. Both clustering approaches follow a similar structure: we first normalize the doculect-by-correspondence tally matrix to adjust feature frequencies by how informative they are, then we perform hierarchical clustering. Each approach is carried out once with only the context-less sound correspondences and once with all context types.

4.3.1 Agglomerative Clustering

This approach is similar to a method used by Prokić et al. (2012) in that it involves agglomerative clustering. However, we use different procedures for measuring distances between pairs of doculects and for transforming the distance values into a hierarchical structure.

We first transform the frequencies in the doculect-by-correspondence tally matrix by applying TF-IDF weighting.

Term frequency (TF) measures the relative frequency of each sound correspondence within a doculect (Luhn, 1957):

$$\mathrm{tf}(doculect_i, corres_j) = \frac{\text{number of occurrences of } corres_j \text{ in } doculect_i}{\text{number of occurrences of all sound correspondences in } doculect_i}$$

while inverse document frequency (IDF) considers how many doculects cover a given sound correspondence (Spärck Jones, 1972):

$$\mathrm{idf}(corres_j) = log(\frac{\text{number of doculects}}{\text{number of doculects with } corres_j}).$$

To combine term frequency and inverse document frequency and transform the tally matrix, we use the implementation from the Python library scikit-learn (Pedregosa et al., 2011), where TF-IDF is calculated as

$$\text{tf-idf}(doculect_i, corres_j) = \text{tf}(doculect_i, corres_j) \quad (\text{idf}(corres_j) + 1).$$

We create a doculect-by-doculect distance matrix with distances bounded between 0 (identical) and 1 (maximally different) by calculating the cosine distances between each binary combination of row vectors from the doculect-by-correspondence matrix (where $doculect_i$ and $doculect_j$ refer to the ith and jth row vectors, respectively):

$$\text{cosine\_distance}(doculect_i, doculect_j) = 1 \quad \frac{doculect_i \quad doculect_j}{\mathsf{k}doculect_i\mathsf{kk}doculect_j\mathsf{k}}.$$

We then convert this distance matrix into a dendrogram using the Unweighted Pair Group Method using Arithmetic Averages (UPGMA) method introduced by Sokal and Michener (1958):

1. Each doculect forms a singleton cluster.
2. The two most similar clusters are merged into a new cluster. The distance between this new cluster B and any given cluster C is

$$\text{dist}(B, C) = \sum_{x2B} \sum_{y2C} \frac{\text{cosine\_distance}(x, y)}{\mathsf{j}B\mathsf{j} \quad \mathsf{j}C\mathsf{j}}.$$

3. Repeat step 2 until only a single cluster containing all doculects is left.

UPGMA was found to be preferable to other distance matrix-based hierarchical clustering methods for analyzing dialect distances by Heeringa (2004, p. 153), and among several clustering methods suited for dialectometry by Prokić and Nerbonne (2008).

Henceforth, we refer to the results of this approach as *UPGMA-context* and *UPGMA-nocontext*.

4.3.2 Bipartite Spectral Graph Co-clustering

For the other clustering method, we use the approach introduced by Dhillon (2001). We follow Wieling and Nerbonne who introduced this method to dialectometry for flat clustering (2009; 2011) and hierarchical clustering (2010).

For this approach, we use a binary version of the doculect-by-correspondence tally matrix that only indicates whether a doculect exhibits a sound correspondence (1) or not (0).

This method works as follows:

1. We begin with normalizing the binary co-occurrence matrix $A \; 2 \; \mathsf{R}^{m \;\; n}$ (m = number of doculects, n = number of sound correspondences). First,

we create two diagonal matrices $D_1 \in \mathbb{R}^{m \times m}$ and $D_2 \in \mathbb{R}^{n \times n}$ that, respectively, contain the row sums and column sums of A. We use these diagonal matrices to reduce the importance of doculects (or correspondences) that co-occur with a large number of sound correspondences (or doculects). Accordingly, we create the normalized matrix A_n by dividing each entry in A by the square root of the sum of its row's entries and by the square root of the sum of its column's entries:

$$A_n = D_1^{\frac{1}{2}} \cdot A \cdot D_2^{\frac{1}{2}}.$$

2. We perform singular value decomposition on A_n, that is, we decompose A_n into the product of three matrices such that $A_n = U\Sigma V^T$ ($U \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{m \times n}_0$ being a diagonal matrix with values in descending order, $V \in \mathbb{R}^{n \times n}$) to obtain the left and right singular vectors u_i and v_i (columns of U and V, respectively). We ignore the singular vectors belonging to the largest singular value as they do not contain information relevant for partitioning the data (Kluger, 2003), and work with the second singular vectors (u_2, v_2) instead. We calculate the vector $Z \in \mathbb{R}^{(m+n) \times 1}$ such that its first m entries contain information about the doculects and the following n entries about the sound correspondences:

$$Z_{[0,m]} = D_1^{\frac{1}{2}} \cdot u_2$$
$$Z_{[m,m+n]} = D_2^{\frac{1}{2}} \cdot v_2.$$

3. We perform k-means clustering on Z with $k = 2$.[14]
4. For each cluster that contains at least two doculects, we create the binary co-occurrence matrix A describing the doculects and sound correspondences in this cluster, and repeat all steps.

If a cluster produced in step 3 contains sound correspondences that are not exhibited by any of the doculects in this cluster, we assign this correspondence to the other cluster. This only happens rarely, and in these cases the corresponding value in Z is near the k-means decision boundary. We need to change the cluster identity in such situations, as it would otherwise not be possible to normalize the cluster's co-occurrence matrix when partitioning the cluster elements again.

The results from this method are hereafter referred to as *UPGMA-context* and *UPGMA-nocontext*.

4.4 Ranking Sound Correspondences by Importance

When all doculects have been assigned to this hierarchical cluster structure, we rank the sound correspondences associated with each cluster. In the case of the UPGMA method, these are all sound correspondences exhibited by each cluster's doculects; for the graph clustering method, these are the sound correspondences that are in the same cluster.

[14] We use the k-means++ algorithm (Arthur and Vassilvitskii, 2007) for semi-arbitrarily picking the initial cluster centres.

We use the representativeness and distinctiveness metrics introduced by Wieling and Nerbonne (2011), as well as a modified version of their importance measure.

Representativeness measures how many doculects in a given cluster exhibit a given sound correspondence:

$$\text{rep}(cluster_i, corres_j) = \frac{\text{number of doculects in } cluster_i \text{ with } corres_j}{\text{number of doculects in } cluster_i}.$$

Representativeness is bounded between 0 (no doculects in the cluster show the given sound correspondence) and 1 (all doculects in the cluster do).

Distinctiveness indicates how often a given sound correspondence occurs in a given cluster compared to other clusters. This requires two additional measures: relative occurrence, which indicates the proportion of doculects exhibiting a given sound correspondence in a given cluster, and relative size, which gives the number of doculects in the cluster relative to the number of all examined modern doculects:

$$\text{relative\_occurrence}(cluster_i, corres_j) = \frac{\text{number of doculects in } cluster_i \text{ with } corres_j}{\text{total number of doculects with } corres_j}$$

$$\text{relative\_size}(cluster_i) = \frac{\text{number of doculects in } cluster_i}{\text{total number of doculects}}.$$

These two concepts are combined to determine the distinctiveness score:

$$\text{dist}(cluster_i, corres_j) = \frac{\text{relative\_occurrence}(cluster_i, corres_j) \quad \text{relative\_size}(cluster_i)}{1 \quad \text{relative\_size}(cluster_i)}.$$

Distinctiveness has an upper bound of 1 (a given sound correspondence only occurs in a given cluster), but no lower bound. A value of 0 means that the sound correspondence has the same relative frequency within the cluster as among the total set of doculects. Negative values indicate that the sound correspondence is (proportionally) rarer within the cluster than among all doculects.

Importance is the average of representativeness and distinctiveness. Wieling and Nerbonne (2011) use the arithmetic mean and mention the possibility of exploring other ways of combining the two metrics. We use the harmonic mean in order to penalize cases where the representativeness value is very high but the distinctiveness value is very low (or vice versa). We also assign an importance score of 0 to cases with negative distinctiveness values:

$$\text{imp}(cluster_i, corres_j) = \begin{cases} \dfrac{2 \quad \text{rep}(cluster_i, corres_j) \quad \text{dist}(cluster_i, corres_j)}{\text{rep}(cluster_i, corres_j) + \text{dist}(cluster_i, corres_j)}, & \text{if dist}(cluster_i, corres_j) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We additionally re-rank correspondences with the same importance score so that more frequent correspondences rank higher.

5 Results

The sound correspondence extraction for our data yields 201 correspondences without context information, 292 with simple context information, 111 with sound class-based context information, and 62 with word boundary information.

Using these sound correspondences for applying the aforementioned methods results in four arrangements of the data into hierarchical partitions, two for the UPGMA method and two for the graph clustering method.

5.1 Agglomerative Clustering

Figure 4 shows the dendrograms created by the UPGMA method for sound correspondences including and excluding contextual information. Of the 18 intermediary clusters (i.e. clusters that are neither singletons nor contain all doculects), 13 are associated with a sound correspondence with an importance score of at least 70% for the context-less run, and 17 for the run with additional contextual information.

The cosine similarity table which is the basis for the UPGMA-context dendrogram is described in section 5.3.3.

In total, 201 sound correspondence were used for the run without contextual information, of which 6 have importance scores of 100% for intermediary clusters. For the run with contextual information, 24 sound correspondences (of 665 total) reach 100% importance for intermediary clusters. Tables 4 and 5 show the highest-ranking sound correspondences (importance score 90%) for UPGMA-nocontext and UPGMA-context, respectively.

Cluster	Sound corres.	Imp.	Rep.	Dist.	Count
Cologne, Luxembourg	x > S	100	100	100	6
Herrlisheim, Ortisei	a > 0	100	100	100	10
Heligoland, Westerkwartier	t > d	100	100	100	8
Ostend, Antwerp, Std. Dutch (BE)	f > v	100	100	100	15
	k > k	100	100	100	13
Hard, Biel, Walser, Graubünden	e > y:	100	100	100	13
Feer, Cologne, Limburg, Luxembourg, Herrlisheim, Tübingen, Hard, Biel, Walser, Graubünden, Ortisei	t > s	90	82	100	59
Feer, Heligoland, Westerkwartier, Cologne, Limburg, Luxembourg, Herrlisheim, Tübingen, Hard, Biel, Walser, Graubünden, Ortisei	; > P	96	92	100	169
	s > S	92	85	100	107

Table 4: Sound correspondences with an importance score of 90% or higher for UPGMA-nocontext. Importance, representativeness, and distinctiveness scores are percentages and rounded to the nearest integer.

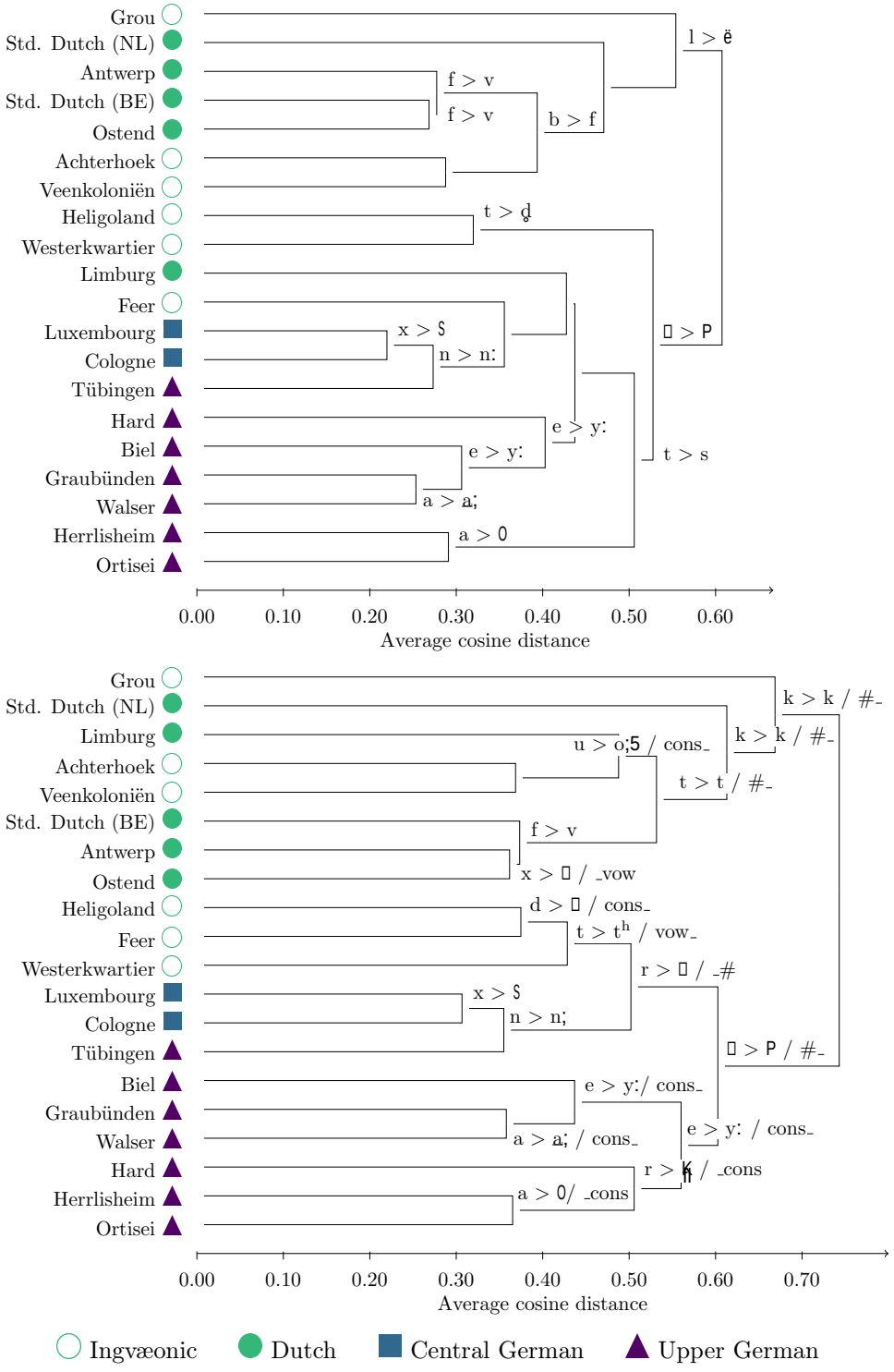

Figure 4: UPGMA with no (top) and additional (bottom) context information, as well as the highest-ranking correspondence per non-singleton cluster (with 70% importance).

Cluster	Sound corres.	Imp.	Rep.	Dist.	Count
Cologne, Luxembourg	x > $	100	100	100	6
	x > $/ vow_	100	100	100	6
Walser, Graubünden	a > a; / cons_	100	100	100	7
Ostend, Antwerp	x > ; / _vow	100	100	100	17
	r > s	100	100	100	11
	r > s / vow_	100	100	100	10
	k > k / vow_	100	100	100	6
Herrlisheim, Ortisei	a > 0	100	100	100	10
	a > 0/ _cons	100	100	100	10
	a > 0/ cons_	100	100	100	9
	r > X/ _cons	100	100	100	9
Ostend, Antwerp, Std. Dutch (BE)	f > v	100	100	100	15
	f > v / _vow	100	100	100	13
	k > k	100	100	100	13
	f > v / #_	100	100	100	12
Feer, Heligoland	d > ; / cons_	100	100	100	10
	d > ; / N_	100	100	100	8
	s > s / _K	100	100	100	8
Feer, Heligoland, Westerkwartier,	; > P	100	100	100	169
Cologne, Luxembourg, Herrlisheim,	; > P / #_	100	100	100	169
Tübingen, Hard, Biel,	; > P / _vow	100	100	100	169
Graubünden, Walser, Ortisei	t > ts	91	83	100	76
	k > kʰ	91	83	100	52
	t > ts / #_	91	83	100	49
	k > kʰ / #_	91	83	100	46
	t > ts / _vow	91	83	100	40
	k > kʰ / _vow	91	83	100	38
	r > ; / _vow	91	83	100	30
	ll > l	91	83	100	30
	ll > l / vow_	91	83	100	30
Veenkoloniën, Grou, Achterhoek,	k > k / #_	100	100	100	40
Std. Dutch (NL), Ostend, Antwerp,	t > t / _#	100	100	100	37
Std. Dutch (BE), Limburg	k > k / _vow	100	100	100	31

Table 5: Sound correspondences with an importance score of 90% or higher for UPGMA-context. Importance, representativeness, and distinctiveness scores are percentages and rounded to the nearest integer.

5.2 Bipartite Spectral Graph Clustering

The results for the graph clustering method are displayed in Figure 5. Again, the doculects are sorted into 18 intermediary clusters. Of these, 6 contain at least one sound correspondence with an importance rating of 70% or above for the context-less run, and 10 for the run with additional context information.

Of the 201 sound correspondences for BSGC-nocontext, 1 has an importance score of 100% for an intermediary cluster (the only sound correspondence with an importance value 90%). For the intermediary clusters of BSGC-context, 2 out of 665 correspondences reach an importance score of 100% and a total of 9 sound correspondences reach at least 90%. Table 6 presents these high-ranking sound correspondences for both BSGC runs.

It should be noted that the results concerning smaller subclusters are (due to the random k-means initialization) not stable across runs, although the overall results have stayed similar.

Cluster	Sound corres.	Imp.	Rep.	Dist.	Count
Antwerp, Std. Dutch (BE)	e > e:	100	100	100	7

Cluster	Sound corres.	Imp.	Rep.	Dist.	Count
Heligoland, Westerkwartier	t > t^h / vow_	100	100	100	14
Antwerp, Std. Dutch (BE)	e > e:	100	100	100	7
Feer, Cologne, Limburg, Luxem-	t > s	90	82	100	59
bourg, Herrlisheim, Tübingen, Hard,	t > s / vow_	90	82	100	58
Biel, Walser, Graubünden, Ortisei	t > s / _#	90	82	100	34
Feer, Heligoland, Westerkwartier,	ˈ > P / _vow	96	92	100	169
Cologne, Limburg, Luxembourg,	s > S	92	85	100	107
Herrlisheim, Tübingen, Hard,	s > S / #_	92	85	100	63
Biel, Walser, Graubünden, Ortisei	s > S / _cons	92	85	100	60

Table 6: Sound correspondences with an importance score of 90% or higher for BSGC-nocontext (top) and BSGC-context (bottom). Importance, representativeness, and distinctiveness scores are percentages and rounded to the nearest integer.

5.3 Comparisons to Continental West Germanic Groupings

For all methods, the first (in the figures: rightmost) split creates one group containing only Dutch, Low German and Frisian doculects, and one group containing a mixture of a few doculects belonging to the aforementioned categories as well as mostly Central or Upper German doculects. We refer to the former group as the Northern cluster and to the latter as the Southern cluster.

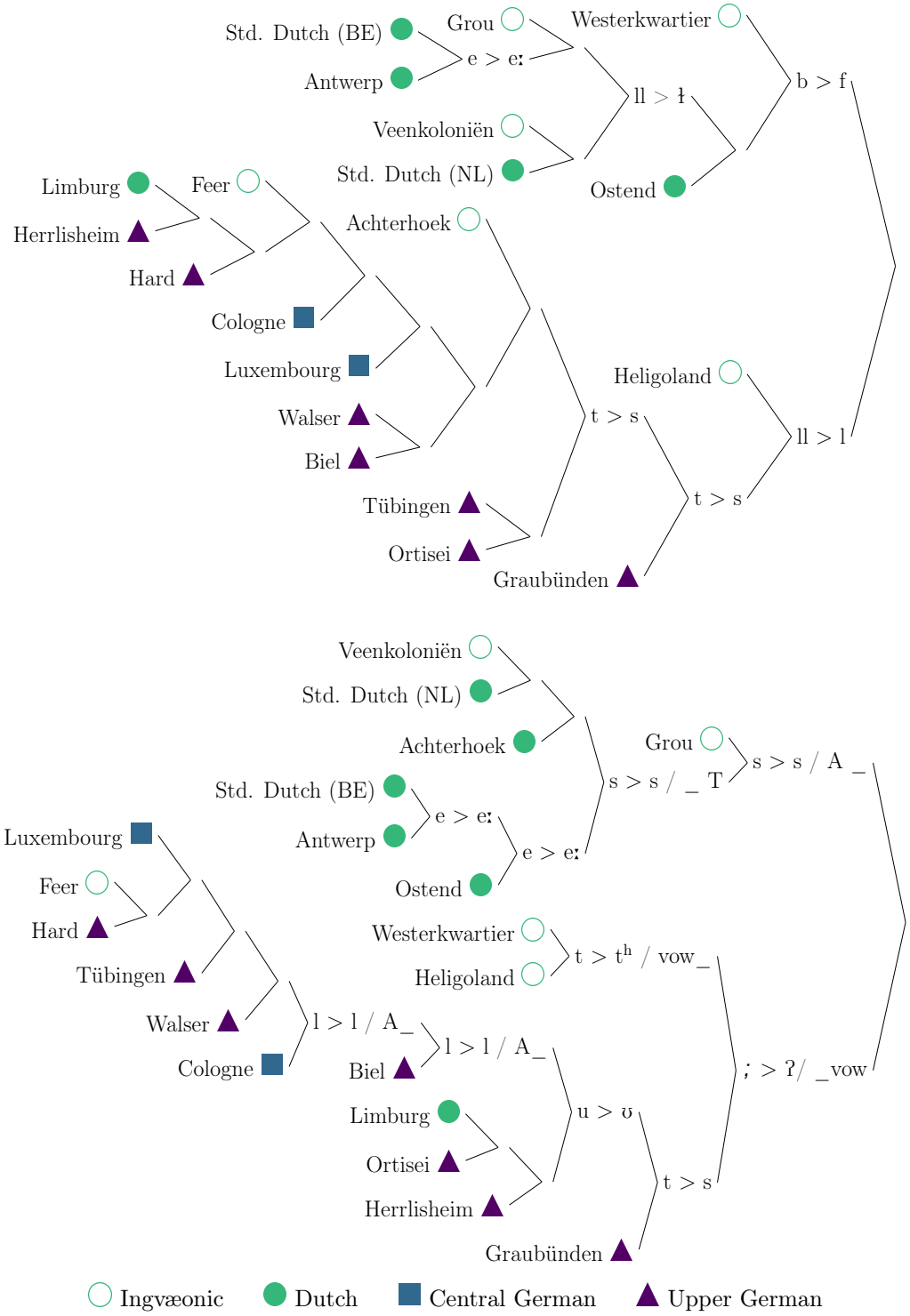

Figure 5: BSGC with no (top) and additional (bottom) context information, as well as the highest-ranking correspondence per non-singleton cluster (with 70% importance).

5.3.1 North Sea Germanic

For all approaches, the Ingvæonic doculects are distributed across several not directly connected clusters. These are, by method, (from least to most connected):

BSGC-nocontext: None of the Ingvæonic doculects share a cluster with only other Ingvæonic doculects.

BSCG-context: Westerkwartier and Heligoland are clustered together.

TFIDF-nocontext: Achterhoek and Veenkoloniën form a cluster, as do Heligoland and Westerkwartier.

TFIDF-context: Achterhoek and Veenkoloniën also form a cluster. Feer, Heligoland and Westerkwartier form another cluster.

All results include several Ingvæonic doculects in the Southern cluster (this is expanded upon in the following subsection). In all cases, these samples include both doculects categorized as *Northern Frisian* in Glottolog (Hammarström et al., 2018): Feer and Heligoland.

The phonological characteristics of Ingvæonic doculects as detailed by Stiles (2013) (changes to /*A/, palatalization of /*k/ and /*g/) are not reflected in any of the results.

5.3.2 Results of the High German Sound Shift

For both BSGC runs, the results can be compared to the consonant shift-based grouping as follows: The majority of the Low German/Dutch/Frisian doculects are in the Northern cluster, although four of them are distributed throughout the Southern cluster. (In BSGC-nocontext, the samples in the latter group are not directly clustered together, but in BSGC-context, two of them are (Westerkwartier and Heligoland).) The Central German doculects from Cologne and Luxembourg are not directly clustered together; the smallest cluster they share is with four other doculects. Of the Upper German doculects, Walser and Biel are directly clustered together, as are Tübingen and Ortisei. The other three Upper German samples are distributed throughout the Southern cluster.

By contrast, the results of the UPGMA runs fit the grouping based on the High German sound shift more closely. Again, most of the samples from the Low German/Dutch/Frisian group constitute the Northern cluster. The remaining samples from this class are clustered together to different degrees in the different runs: In UPGMA-nocontext, Heligoland and Westerkwartier are directly grouped together, but the remaining two doculects (Feer and Limburg) are not. In UPGMA-context on the other hand, there are only three doculects outside the Northern cluster (Heligoland, Feer, Westerkwartier), and they are all grouped together.

Both Central German doculects are directly clustered together for both UPGMA runs. In addition, for both runs, one of the Upper German doculects (Tübingen) is grouped with the Central German samples; the rest form their own cluster.

The outcomes of the High German sound shift or lack thereof are also visible in some of the highest-ranked sound correspondences.

In case of the UPGMA-context run, the sound correspondence rules with 100% importance scores for (subclusters within) the Northern cluster all reflect some of the unvoiced stops not being weakened: /*k/ > [k] / #_, /*t/ > [t] / _#, /*k/ > [k] / _*vow*.

Conversely, the predominantly Upper German clusters have high-ranked correspondences demonstrating lenition: /*t/ > [t͡s], /*t/ > [t͡s] / #_, /*t/ > [t͡s] / _*vow* (UPGMA-context) and /*t/ > [s] (UPGMA-nocontext, BSGC). None of the sound correspondences with high importance values describe the affrication or spirantization of /*k/ or /*p/ for any of the clustering approaches, although it has to be noted that with the given data, none of the sound correspondences based on /*p/ occurred often enough to actually be used for clustering.

Nevertheless, the sound correspondence /*k/ > [kʰ] (on its own or prevocally) also has high importance scores for the Southern group. However, this still matches the assertion that /*k/ was the most resistant to change during the High German sound shift.

5.3.3 Close Doculects Outside these Groupings

In this section, we consider the results from the UPGMA-context run more closely, since it is the method whose results match the groupings in the literature the most. The cosine similarity matrix that is the basis for the hierarchical clustering step is visualized in Figure 6.

The pairwise cosine similarity values showcased in this figure show geographically close as well as geographically distant connections with high similarity scores, some of which are less apparent after the hierarchical clustering step.

Proceeding from (North) West to (South) East on the map, we get the following connections:

Among the doculects spoken in Belgium and the Netherlands, there is a tight cluster of Belgian doculects (Ostend, Antwerp, Std. Dutch (BE)) which are connected to the Veenkoloniën doculect via Ostend, which is in turn very similar to that spoken in Achterhoek.

The doculect spoken in Westerkwartier is very dissimilar to the doculects which are its direct neighbours (Grou and Veenkoloniën), but it forms a tight cluster with Heligoland and Feer.

Heligoland and Feer are also very similar to the Luxembourg doculect. Luxembourg, Feer and Tübingen are directly connected, as are Luxembourg, Tübingen and Cologne. However, there are low cosine similarity scores between the latter three doculects and the doculects from Limburg and Herrlisheim, both of which are geographically close to this group.

Tübingen, Graubünden and Walser form another triangle cluster with high pairwise cosine similarity scores. Walser is also very similar to the Hard doculect,

which is only moderately similar to the samples from Tübingen and Graubünden.

Lastly, the doculects spoken in Herrlisheim and Ortisei show high cosine similarity to one another but not to their geographic neighbours.

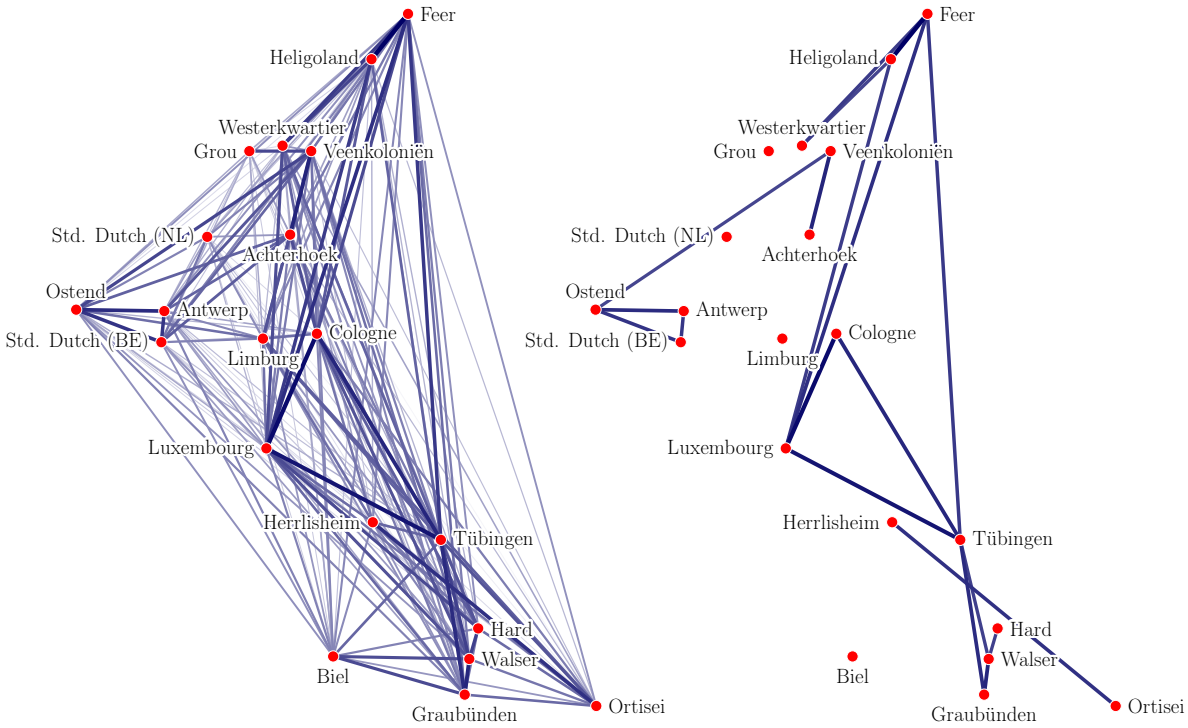

Figure 6: Cosine similarities between the doculects (UPGMA-context). Lines that are bolder and darker represent greater cosine similarity scores (i.e. lower cosine distances). The graphic on the left includes all pairwise similarity scores; the one on the right only includes the highest 10% of cosine similarity scores.

5.4 Context Information

As shown in the preceding subsections, the runs with context information yield results that are overall closer to the proposed groupings than the runs without additional information.

Moreover, for the runs with additional context information, higher proportions of the clusters are associated with sound correspondences (of 70% importance): 6/18 versus 10/18 for the graph clustering approach (without and with context information, respectively), and 13/18 versus 17/18 for UPGMA-nocontext and UPGMA-context.

Most of the high-ranking sound correspondences use the vowel/consonant distinction instead of the more detailed sound classes. Additionally, word boundary-

based correspondences are also very common among the high-ranking sound correspondences.

6 Discussion

6.1 Clusters

The clusters formed by the different approaches match the groupings from the literature to different degrees. The UPGMA runs show these patterns more strongly than the BSGC runs, and UPGMA-context follows it the most closely.

However, even in UPGMA-context, we see some unexpected results, such as several Frisian/Low German doculects being clustered with a group of Central/Upper German doculects instead of the large Frisian/Low German/Dutch group.

Overall, we can also observe that the results reflect the High German sound shift-based groupings more strongly than the (Non-)Ingvæonic distinction, unless we (re-)define Ingvæonic to include Dutch as well, as Sonderegger (1979) does.

Inspecting the cosine similarity values for UPGMA-context, we can observe a pattern that shows both some isolated subgroups and a more net-like pattern of connections reflecting more gradual changes. These latter pattern matches the observations by Heggarty et al. (2010) for the same dataset. It would be interesting to investigate these gradual changes further, for instance following the fuzzy dialect clustering approaches by Pröll (2013), which are better suited to model wave-like developments of language variation.

A promising approach would be to run clustering methods (fuzzy or not) on a larger amount of data. This should be both in terms of the number of doculects to gain a more representative depiction of the CWG doculect landscape (adding places that are geographically located in between some of the doculects we worked with, and adding doculects from CWG-speaking places that were not at all included in this analysis, such as more locations in Germany)[15] and in terms of concepts per doculect (to capture a greater variety of sound correspondences). Unfortunately, we are not aware of a digital database compiled by a single transcriber that would allow us to do that.

6.2 Bipartite Spectral Graph Co-clustering

As mentioned before, the co-clustering approach does not match the groupings described in the literature very well compared to the simpler UPGMA approach.

In our experiments, one issue with BSGC is that some sound correspondences get assigned to the wrong clusters in that they do not actually co-occur with any of the doculects in their assigned clusters. In such cases, we automatically

[15] Adding more doculects spoken near the Dutch border would make it possible to investigate the tendency for standard language borders to act as dialect borders. In addition, incorporating data from central or eastern Germany would noticeably increase the diversity of CWG doculects covered in such an investigation.

assign them to the appropriate other cluster to avoid linear algebra problems. However, it is possible that such an unfitting cluster assignment also happens with other samples that get sorted into the cluster whose doculects they not describe as well as the other cluster's.

In the future, it would be interesting to explore if alternate approaches to normalizing the data would mitigate this problem.

However, BSGC yielded good results for Wieling and Nerbonne (2009; 2010; 2011), Wieling et al. (2013) and Montemagni et al. (2013). There are several differences between their data and ours. First, all of these experiments used a larger number of doculects and concepts. Additionally, the doculects they used are generally from smaller geographic areas. Moreover, all of them used modern reference dialects. The transcriptions they worked with might also have been broader, potentially allowing for sound correspondences that are similar in our data, but (due to the narrowness of transcription) slightly different, to become identical and thus more common.

Another follow-up investigation would be to investigate the influence of data selection and preprocessing on BSGC performance.

6.3 Sound Correspondences

Adding context information to the sound correspondences helps the UPGMA model to match the groupings in the literature more closely. However, most of the high-ranking sound correspondences do not involve the more detailed sound class model. It is possible that these sound classes are too specific for the data at hand or have the wrong kind of specificity, being thus not well-suited for capturing generalizations of the data.

A solution to this would be using a more simple context system. Additionally, it would also be possible to use more sound segment and context representations with varying degrees of superficiality or abstractness to model sound changes as specifically or generally as appropriate for each resulting cluster.

The current system cannot pick up on related changes with slightly different outcomes, such as different ways that Proto-Germanic /*k/ has developed into an affricate or fricative in Swiss German doculects: /*k/ > [kx͡x], /*k/ > [x], /*k/ > [X]. Again, a way of describing these sound changes more abstractly might make it possible to consider the similarity of these sound changes while clustering doculects.

In future experiments, it might be worthwhile to rank sound correspondences also by how regular they are, i.e. how often a given sound shifted into a specific sound compared to how often another sound shift (or no sound shift) took place instead. This could be done in a similar fashion to the approaches by Prokić (2007) and Prokić and Cysouw (2013).

In this thesis, all of the comparisons between Proto-Germanic and modern CWG doculects are phonological and on a word level. Investigating the effect of additionally incorporating morphological, lexical and syntactical information would be interesting for further research.

7 Conclusion

In this thesis, we have implemented two methods for clustering doculects on the basis of shared historical sound correspondences, and compared the results to common (although not uncontroversial) groupings of the doculects we worked with.

We showed that, compared to the graph co-clustering method (BSGC), the agglomerative clustering approach (UPGMA) yielded results that are more similar to the groupings in the literature and are associated more frequently sound correspondences that describe the subclusters well.

Additionally, we examined the effect of adding information about the phonetic context in which specific sound shifts take place, and showed that adding such information also resulted in clusters that match the literature more closely and that can more frequently be described with relevant sound correspondences.

Further investigation is needed for exploring the differences in the results between the two clustering approaches. Moreover, we hope to examine how the representation of sound shifts can be improved to describe them in as much detail or abstraction as required for good clustering results. In addition, it would be worthwhile to pursue ways of combining these approaches with fuzzy clustering techniques to better capture transitions within dialect continua.

References

Arthur, D. and S. Vassilvitskii (2007). k-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics.

Bryant, D. and V. Moulton (2004). Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution 21*(2), 255–265.

Chomsky, N. and M. Halle (1968). *The Sound Pattern of English*. Harper & Row.

Dhillon, I. S. (2001). Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, San Francisco, pp. 269–274. ACM.

Goblirsch, K. G. (2005). *Lautverschiebungen in den germanischen Sprachen*. Heidelberg: Winter.

Hammarström, H., R. Forkel, and M. Haspelmath (2018). Glottolog 3.3. Max Planck Institute for the Science of Human History.

Harbert, W. (2007). *The Germanic Languages*. Cambridge University Press.

Heeringa, W. J. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph. D. thesis, Citeseer.

Heggarty, P. (2018). Sound Comparisons: Exploring Diversity in Phonetics across Language Families. www.soundcomparisons.com.

Heggarty, P., W. Maguire, and A. McMahon (2010). Splits or Waves? Trees or Webs? How Divergence Measures and Network Analysis can Unravel Language Histories. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 365*(1559), 3829–3843.

Jones, E., T. Oliphant, P. Peterson, et al. (2001). SciPy: Open source scientific tools for Python.

Kluger, Y. (2003). Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome Research 13*(4), 703–716.

König, W. (2005). *DTV-Atlas Deutsche Sprache* (18 ed.). Deutscher Taschenbuch Verlag. Graphics by Hans-Joachim Paul.

Kremer, L. and H. Niebaum (1990). Zur Einführung: Grenzdialekte als Gradmesser des Sprachwandels. In L. Kremer and H. Niebaum (Eds.), *Grenzdialekte. Studien zur Entwicklung kontinentalwestgermanischer Dialektkontinua*, pp. 7–21. Olms.

List, J.-M. (2012). SCA: Phonetic Alignment Based on Sound Classes. In M. Slavkovik and D. Lassiter (Eds.), *New Directions in Logic, Language and Computation*, pp. 32–51. Berlin and Heidelberg: Springer.

List, J.-M., S. Greenhill, and R. Forkel (2018). LingPy: A Python Library for Historical Linguistics. Version 2.6.3. With contributions by Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel, and Tiago Tresoldi.

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development 1*(4), 309–317.

Montemagni, S., M. Wieling, B. de Jonge, and J. Nerbonne (2013). Synchronic Patterns of Tuscan Phonetic Variation and Diachronic Change: Evidence from a Dialectometric Study. *Literary and Linguistic Computing 28*(1), 157–172.

Needleman, S. B. and C. D. Wunsch (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology 48*(3), 443–453.

Nerbonne, J. (2009). Data-driven Dialectology. *Language and Linguistics Compass 3*(1), 175–198.

Nielsen, H. F. (1989). *The Germanic Languages: Origins and Early Dialectal Interrelations*. The University of Alabama Press.

Noble, C. A. M. (1983). *Modern German Dialects*, Volume 519 of *Europäische Hochschulschriften. Reihe 1, Deutsche Sprache und Literatur*. Peter Lang.

Notredame, C., D. G. Higgins, and J. Heringa (2000). T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *Journal of Molecular Biology 302*(1), 205–217.

Oliphant, T. E. (2006). *A Guide to NumPy*. Trelgol Publishing USA.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Prokić, J. (2007). Identifying Linguistic Structure in a Quantitative Analysis of Dialect Pronunciation. In *Proceedings of the ACL 2007 Student Research Workshop*, pp. 61–66. Association for Computational Linguistics.

Prokić, J., Ç. Çöltekin, and J. Nerbonne (2012). Detecting Shibboleths. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pp. 72–80. Association for Computational Linguistics.

Prokić, J. and M. Cysouw (2013). Combining Regular Sound Correspondences and Geographic Spread. *Language Dynamics and Change 3*(2), 147–168.

Prokić, J. and J. Nerbonne (2008). Recognizing Groups among Dialects. *International Journal of Humanities and Arts Computing 2*(1-2), 153–172.

Pröll, S. (2013). Detecting Structures in Linguistic Maps: Fuzzy Clustering for Pattern Recognition in Geostatistical Dialectometry. *Literary and Linguistic Computing 28*(1), 108–118.

Renfrew, C. and P. Heggarty (2009). Languages and Origins in Europe. www.languagesandpeoples.com.

Ringe, D. R. (2012). Cladistic Principles and Linguistic reality: The case of West Germanic. In P. Probert and A. Willi (Eds.), *Laws and Rules on Indo European*, pp. 33–42. Oxford University Press.

Sokal, R. R. and C. D. Michener (1958). A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin 28*, 1409–1438.

Sonderegger, S. (1979). *Einführung, Genealogie, Konstanten*, Volume 1 of *Grundzüge deutscher Sprachgeschichte. Diachronie des Sprachsystems*. Walter de Gruyter.

Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation 28*(1), 11–21.

Stiles, P. V. (2013). The Pan-West Germanic Isoglosses and the Subrelationships of West Germanic to Other Branches. *NOWELE. North-Western European Language Evolution 66*(1), 5–38.

Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994). CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research 22*, 4673–4680.

van Coetsem, F. (1992). The Interaction between Dialect and Standard Language, Viewed from the Standpoint of the Germanic Languages. In J. A. van Leuvensteijn (Ed.), *Dialect and Standard Language in the English, Dutch, German and Norwegian Language Areas [Proceedings of the Colloquium "Dialect and Standard Language", Amsterdam, 15–18 October 1990]*. North-Holland.

van der Auwera, J. and D. Van Olmen (2017). The Germanic Languages and Areal Linguistics. In R. Hickey (Ed.), *The Cambridge Handbook of Areal Linguistics*, pp. 239–269. Cambridge University Press.

Voyles, J. B. (1971). The Problem of West Germanic. *Folia Linguistica 5*, 117–150.

Wettig, H., K. Reshetnikov, and R. Yangarber (2012). Using Context and Phonetic Features in Models of Etymological Sound Change. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pp. 108–116. Association for Computational Linguistics.

Wieling, M. and J. Nerbonne (2009). Bipartite Spectral Graph Partitioning to Co-cluster Varieties and Sound Correspondences in Dialectology. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pp. 14–22. Association for Computational Linguistics.

Wieling, M. and J. Nerbonne (2010). Hierarchical Spectral Partitioning of Bipartite Graphs to Cluster Dialects and Identify Distinguishing Features. In *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, pp. 33–41. Association for Computational Linguistics.

Wieling, M. and J. Nerbonne (2011). Bipartite Spectral Graph Partitioning for Clustering Dialect Varieties and Detecting their Linguistic Features. *Computer Speech & Language 25*(3), 700–715.

Wieling, M. and J. Nerbonne (2015). Advances in Dialectometry. *Annual Review of Linguistics 1*, 243–264.

Wieling, M., R. G. Shackleton Jr., and J. Nerbonne (2013). Analyzing Phonetic Variation in the Traditional English Dialects: Simultaneously Clustering Dialects and Phonetic Features. *Literary and Linguistic Computing 28*(1), 31–41.

Zha, H., X. He, C. Ding, H. Simon, and M. Gu (2001). Bipartite Graph Partitioning and Data Clustering. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pp. 25–32. ACM.