Summary: Clustering Dialect Varieties Based on Historical Sound Correspondences (Bachelor's Thesis)

Author: Verena Blaschke, Supervisor: Dr. Çağrı Çöltekin

1 Introduction

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.

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. 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. In doing so, we examine the performance of two different clustering algorithms.

2 Continental West Germanic

The continental European West Germanic (henceforth: CWG) language varieties (hereafter referred to as *doculects*) include several standard languages (Luxembourgish, multiple standard varieties of Dutch and German) as well as many regiolects and dialects. Establishing subgroups within this collection of doculects provides a challenge due to their being very similar to one another.

Nevertheless, there are some common ways of grouping CWG doculects into broad sub-groups. Firstly, they can be divided into Ingvæonic (Low German and Frisian) and non-Ingvæonic varieties, based on pronoun systems and morphological and phonological characteristics (Stiles (2013); Harbert (2007, pp. 7–8)).

Secondly, CWG doculects can be divided into three groups according to the results of the High German sound shift, a lenition of voiceless (Proto-)Germanic stops to affricates or fricatives: Upper German (which almost completely exhibits lenition for all voiceless stops), Central German (which shows a partial development of the High German sound shift), and Low German as well as Dutch and Frisian (which were not influenced by the High German sound shift) (Noble (1983, pp. 33, 55); König (2005, pp. 64, 230–231)).

2.1 Data

We work with phonetically transcribed data from CWG doculects, taken from the Sound Comparisons project (Heggarty, 2018). We used 110 cognate sets (also referred to as *concepts*) from 20 modern CWG doculects and a reconstructed version of Proto-Germanic.

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 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 the modern CWG doculects.

3 Methods

We first align the phonetic transcriptions from our data to then extract sound correspondences, which we use for two different clustering methods.

3.1 Multiple Sequence Alignment

We carry out alignment based on data from all the investigated doculects at once using multiple sequence alignment. We use a library-based version (Notredame et al., 2000) of the progressive multiple sequence alignment method (Thompson et al., 1994), as implemented in the LingPy library for Python (List et al., 2018).

3.2 Sound Correspondence Extraction

Next, 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. For the latter, firstly, we separately add information about the left and right single-segment context, stating whether it 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. Additionally, we repeat the same process, but use the more fine-grained sound class system by List (2012), which recognizes fifteen consonant and six vowel group. Lastly, we also use sound correspondences that explicitly capture correspondences at word-initial and -final positions.

We ignore gap-gap alignments, and treat cases of segment swaps (metathesis) as normal insertions or deletions. 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 per doculect.

3.3 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.

3.3.1 Agglomerative Clustering

We first normalize the frequencies in the doculect-by-correspondence tally matrix by applying TF-IDF weighting. We calculate the cosine distances between each binary combination of row vectors from this matrix to create a doculect-by-doculect distance matrix.

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), which was found to be well-suited for analyzing dialect distances by Heeringa (2004, p. 153) and Prokić and Nerbonne (2008).

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

3.3.2 Bipartite Spectral Graph Co-clustering

For the other clustering method, we perform Bipartite Spectral Graph Co-clustering (BSGC) (Dhillon, 2001), which was introduced to dialectometry by Wieling and Nerbonne (2009; 2010).

BSGC works by normalizing the doculect-by-correspondence matrix such that all doculects and sound correspondences can be mapped to vectors in the same vector space. Using k-means clustering with $k = 2$, each vector is assigned to one of two clusters. For each new cluster, both the normalization and the clustering step are recursively repeated.

Occasionally, it happens that a cluster produced by the k-means clustering step contains a sound correspondence that is not exhibited by any of the doculects in the same cluster. In that case, it is not possible to normalize the new cluster's co-occurence matrix, and we therefore need to assign the sound correspondence to the other cluster that was produced in the k-means clustering step.

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

3.4 Ranking Sound Correspondences by Importance

When all doculects have been assigned to a hierarchical cluster structure, we rank the sound correspondences associated with each cluster. These are either all sound correspondences exhibited by a cluster's doculects (UPGMA), or the sound correspondences that are in the same cluster (BSGC).

We use the *representativeness* (What proportion of doculects in a given cluster exhibit a given sound correspondence?) and *distinctiveness* (How often does a given sound correspondence occur in a given cluster compared to other clusters?) metrics introduced by Wieling and Nerbonne (2011), as well as a modified version of their *importance* metric (the harmonic mean of representativeness and distinctiveness). For all metrics, higher values are better; the best possible score is 100%.

4 Results

Figures 2 and 3 visualize the results of UPGMA and BSGC, respectively.

One notable difference between the results of the two methods is the number of sound correspondences with importance scores of 100% associated with intermediary clusters (i.e. clusters that are neither singletons nor contain all doculects): More sound correspondences associated with UPGMA have perfect importance scores compared to the results of BSGC (UPGMA-nocontext: 6/201 vs. BSGC-nocontext: 1/201; UPGMA-context: 24/665 vs. BSGC-context: 2/665).

Similarly, most (13/18; UPGMA-nocontext) or almost all (17/18; UPGMA-context) intermediary clusters belonging to the agglomerative clustering approach are associated with a sound correspondence with an importance score of at least 70%, the same is only true for considerably fewer intermediary clusters generated by BSGC (BSGC-nocontext: 6/18, BSGC-context: 10/18).

This also shows that adding phonetic context information to the sound correspondences yields clusters that are more frequently associated with representative and distinctive sound correspondences. 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.

4.1 Comparisons to Continental West Germanic Groupings

For all methods, the first (in the figures: rightmost) split into clusters 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.

For all approaches, the Ingvæonic doculects are distributed across several not directly connected clusters. They are the most isolated from one another in the results of BSGC-nocontext, and the most closely connected to one another for UPGMA-context. The phonological characteristics of Ingvæonic doculects are not reflected in any result's set of high-ranking sound correspondences.

The doculect classification according to the High German sound shift is also reflected more closely by the UPGMA runs than by the results of BSGC. 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. The two Central German doculects are not directly clustered together, and the Upper German doculects 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 more tightly within the Southern group in the UPGMA-context run than in the UPGMA-nocontext run. For both UPGMA runs, both Central German doculects are directly clustered together and one of the Upper German doculects is grouped with them; 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, which describe the lenition or the preservation of Proto-Germanic unvoiced stops. This is most strongly the case for the UPGMA-context run.

Since the results of UPGMA-context match the groupings in the literature the most, we also inspected the corresponding cosine similarity matrix that is the basis for the hierarchical clustering step more closely. It is visualized in Figure 4. The pairwise cosine similarity values showcased in this figure do not only show geographically close, but also geographically distant connections with high similarity scores. This displays a more net-like connection between *all* doculects than apparent after the hierarchical clustering step.

5 Discussion

5.1 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. 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 redefine Ingvæonic to include Dutch as well, as e.g. 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. This latter pattern matches the observations by Heggarty et al. (2010) for the same dataset. To further investigate these gradual changes, we could for instance follow the fuzzy dialect clustering approaches by Pröll (2013).

5.2 Bipartite Spectral Graph Co-clustering

The co-clustering approach does not match the groupings described in the literature very well compared to the simpler UPGMA approach.

However, BSGC yielded good results in dialectological experiments by Wieling and Nerbonne (2009; 2010; 2011), Wieling et al. (2013) and Montemagni et al. (2013). There are several differences between their data and ours: the other experiments used a larger number of concepts and doculects, which also come from smaller geographic areas. Moreover, all of them used modern reference dialects. The transcriptions they worked with might also have been broader than the very narrow transcriptions that we worked with. A possible follow-up investigation would be to investigate the influence of data selection and preprocessing on BSGC performance.

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

5.3 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 have the wrong kind of specificity for capturing generalizations of the data at hand. Further research is needed to determine a suitably specific context system.

In future experiments, it might be worthwhile to also rank sound correspondences by how regular they are, in a similar fashion to the approaches by Prokić (2007) and Prokić and Cysouw (2013).

6 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 more frequently with associated 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.

Appendix: Figures

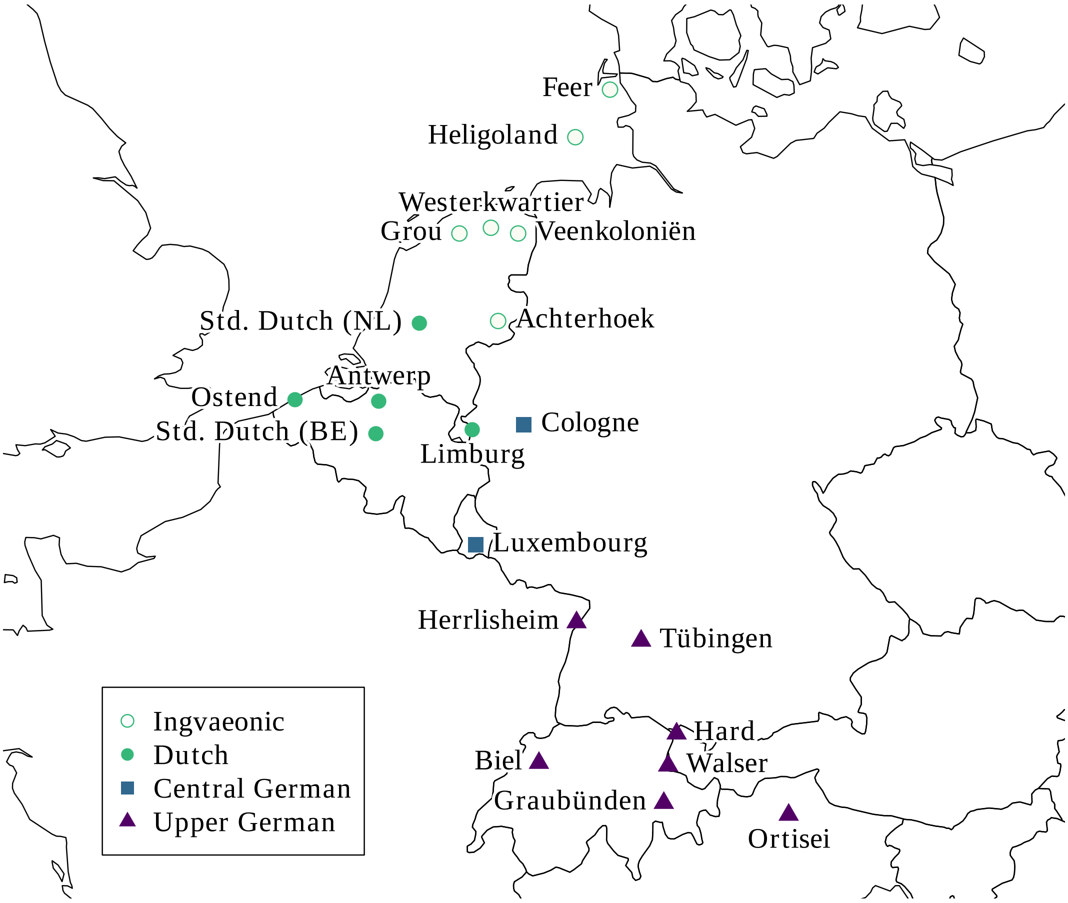

Figure 1: Locations of the modern continental West Germanic doculects we worked with.

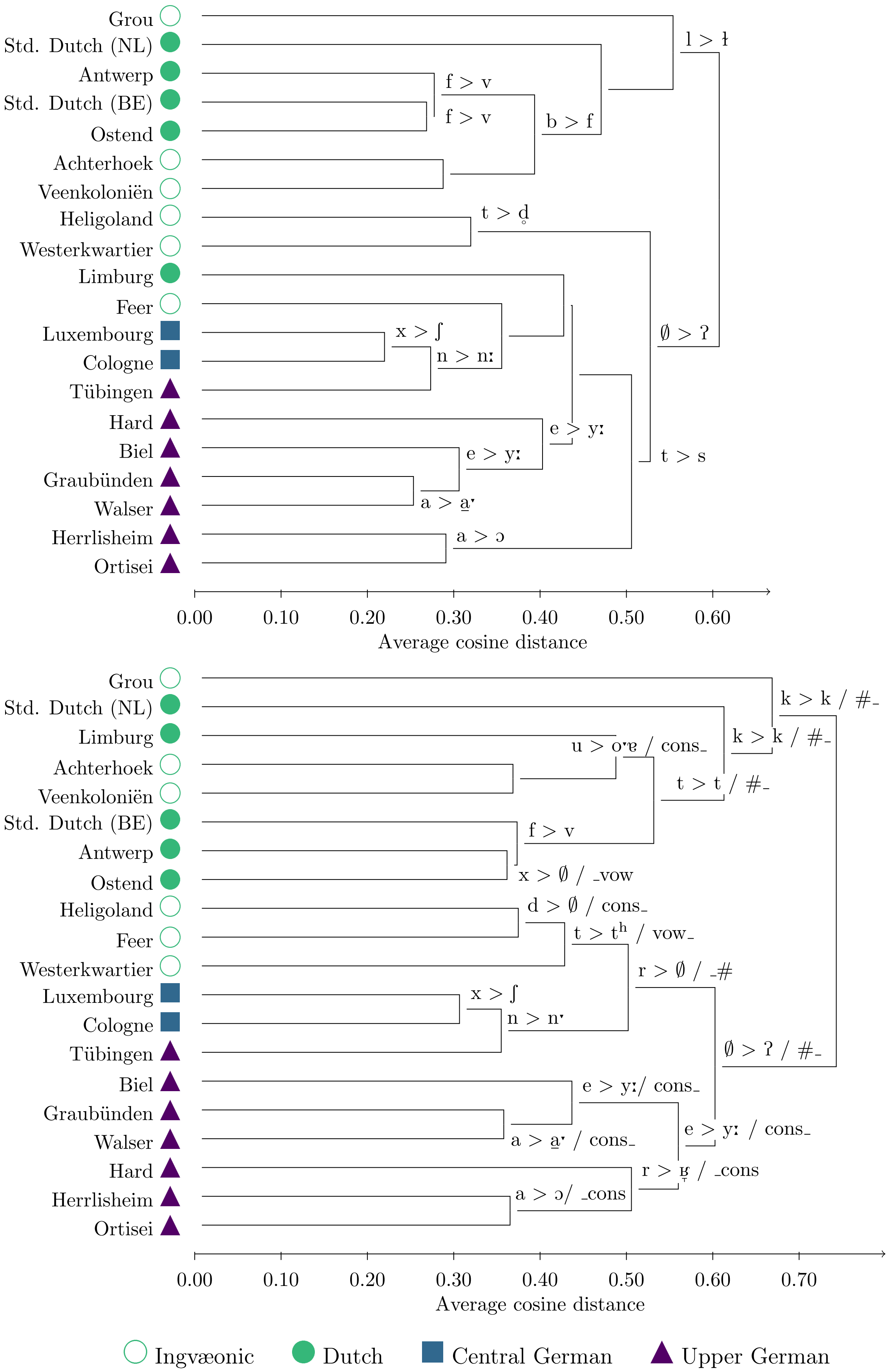

Figure 2: UPGMA with no (top) and additional (bottom) context information, as well as the highest-ranking correspondence per non-singleton cluster (with ≥70% importance). The sound correspondence rules are explained more thoroughly in the full thesis.

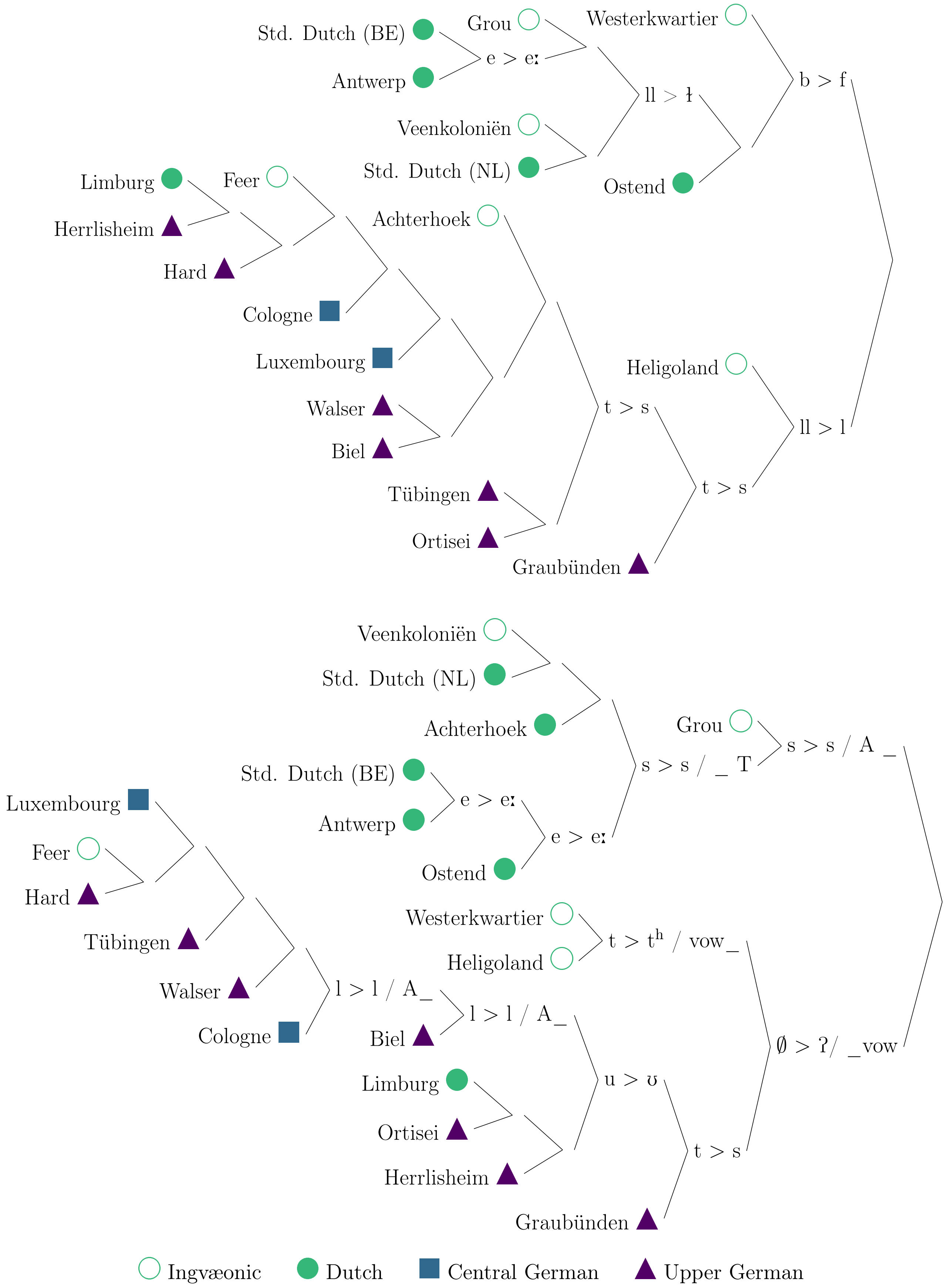

Figure 3: BSGC with no (top) and additional (bottom) context information, as well as the highest-ranking correspondence per non-singleton cluster (with ≥70% importance). The sound correspondence rules are explained more thoroughly in the full thesis.

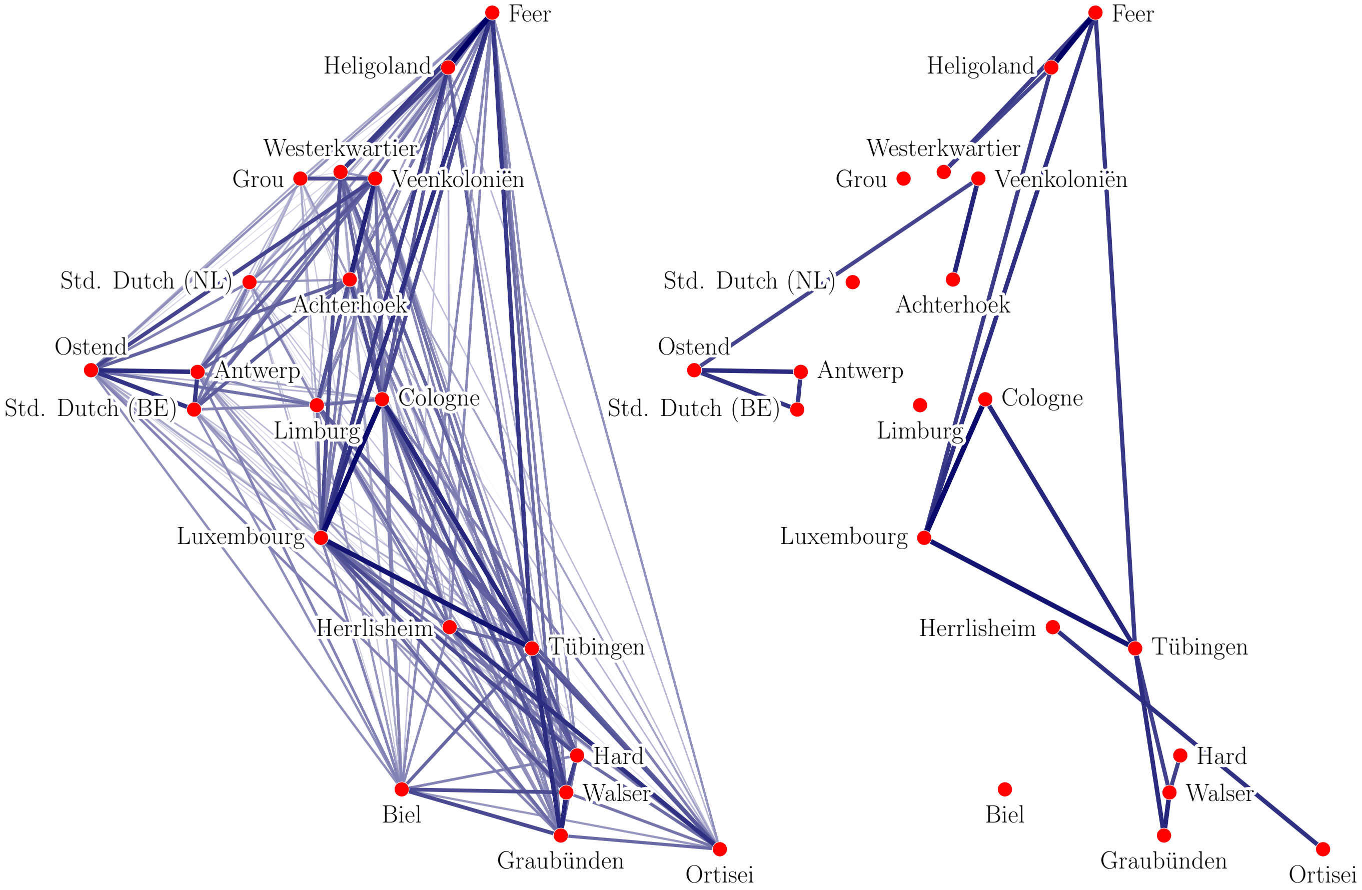

Figure 4: Cosine similarities between the doculects (UPGMA-context). Lines that are bolder and darker represent greater cosine similarity scores (i.e. lower cosine distances). The graphic on the left includes all pairwise similarity scores; the one on the right only includes the highest 10% of cosine similarity scores.

References

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.

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` (Retrieved 2018-05-18).

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.

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

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. `http://lingpy.org`. Version 2.6.3. With contributions by Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel, and Tiago Tresoldi.

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.

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.

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. 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.

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.

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.

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., 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.