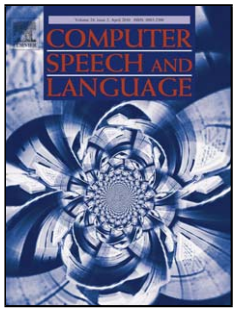
Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features

Martijn Wieling, John Nerbonne



Title: Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features

Authors: Martijn Wieling, John Nerbonne

PII: S0885-2308(10)00041-0
DOI: doi:10.1016/j.csl.2010.05.004
Reference: YCSLA 457

To appear in:

Received date: 13-10-2009
Revised date: 12-2-2010
Accepted date: 7-5-2010

Please cite this article as: Wieling, M., Nerbonne, J., Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features, *Computer Speech & Language* (2008), doi:10.1016/j.csl.2010.05.004

Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features

Martijn Wieling*, John Nerbonne

University of Groningen
Department of Computational Linguistics
Postal addres: Postbus 716, 9700 AS Groningen, The Netherlands
Visiting address: Oude Kijk in ‘t Jatstraat 26, 9712 EK Groningen, The Netherlands
E-mail: m.b.wieling@rug.nl / j.nerbonne@rug.nl
Phone: +31(0)50 363 5977 / +31(0)50 363 5815
Fax: +31(0)50 363 6855

* Corresponding author

Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features☆

Martijn Wieling*,[a], John Nerbonne[a]

[a] *University of Groningen, P.O. Box 716, 9700 AS Groningen, The Netherlands. Tel.: +31(0)50 363 5977. Fax.: +31(0)50 363 6855*

Abstract

In this study we use bipartite spectral graph partitioning to simultaneously cluster varieties and identify their most distinctive linguistic features in Dutch dialect data. While clustering geographical varieties with respect to their features, e.g. pronunciation, is not new, the simultaneous identification of the features which give rise to the geographical clustering presents novel opportunities in dialectometry. Earlier methods aggregated sound differences and clustered on the basis of aggregate differences. The determination of the significant features which co-vary with cluster membership was carried out on a *post hoc* basis. Bipartite spectral graph clustering simultaneously seeks groups of individual features which are strongly associated, even while seeking groups of sites which share subsets of these same features. We show that the application of this method results in clear and sensible geographical groupings and discuss and analyze the importance of the concomitant features.

Key words: Bipartite spectral graph partitioning, Clustering, Sound correspondences, Dialectometry, Dialectology, Language variation

☆This paper is an extended version of the study 'Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology' by Martijn Wieling and John Nerbonne [1]

*Corresponding author

Email addresses: m.b.wieling@rug.nl (Martijn Wieling), j.nerbonne@rug.nl (John Nerbonne)

1. Introduction

Dialect atlases contain a wealth of material that is suitable for the study of the cognitive, but especially the social dynamics of language. Although the material is typically presented cartographically, we may conceptualize it as a large table, where the rows are the sampling sites of the dialect survey and the columns are the linguistic features probed at each site. We inspect a table created for the purpose of illustrating the sort of information we wish to analyze. We used constructed data to keep the point maximally simple.

	HOE?	EEL	...	night	hog	...	p^h	t^h	asp.	...
Appleton	+	A	...	[naɪt]	[hɔg]	...	+	+	+	...
Brownsville	+	A	...	[nat]	[hɑg]	...	+	+	+	...
Charleston	+	B	...	[nat]	[hɑg, hɔg]	...	-	-	-	...
Downe	-	B	...	[nat]	[hɑg]	...	-	-	-	...
Evanston	?	B	...	[naɪt]	[hɔg]	...	+	+	+	...

The features in the first two columns are intended to refer to the cognates, frequently invoked in historical linguistics. The first three varieties (dialects) all have lexicalizations for the concept 'hoe' (a gardening instrument), the fourth does not, and the question does not have a clear answer in the case of the fifth. The first two varieties use the same cognate for the concept 'eel', as do the last three, although the two cognates are different. More detailed material is also collected, e.g. the pronunciations of common words, shown above in the fourth and fifth columns, and our work has primarily aimed at extracting common patterns from such transcriptions. As a closer inspection will reveal, the vowels in the two words suggest geographical conditioning. This illustrates the primary interest in dialect atlas collections: they constitute the empirical basis for demonstrating how geography

influences linguistic variation. On reflection, the influential factor is supposed to be not geography or proximity *simpliciter*, but rather the social contact which geographical proximity facilitates. Assuming that this reflection is correct, the atlas databases provide us with insights into the social dynamics reflected in language.

More abstract characteristics such as whether initial fortis consonants like [p,t] are aspirated (to be realized then as [p^h,t^h]) is sometimes recorded, or, alternatively, the information may be extracted automatically (see references below). Note, however, that we encounter here two variables, aspiration in /p/ and aspiration in /t/, which are strongly associated irrespective of geography or social dynamics. In fact, in all languages which distinguish fortis and lenis plosives /p,b/, /t,d/, etc., it turns out that aspiration is invariably found on *all* (initial) fortis plosives (in stressed syllables), or on none at all, regardless of social conditioning. We thus never find a situation in which /p/ is realized as aspirated ([p^h]) and /t/ as unaspirated. This is exactly the sort of circumstance for which cognitive explanations are generally proposed, i.e. explanations which do not rely on social dynamics. The work we discuss below does not detect or attempt to explain cognitive dynamics in language variation, but the data sets we used should ultimately be analyzed with an eye to cognitive conditioning as well.[1] The present paper focuses exclusively on the social dynamics of variation.

Exact methods have been applied successfully to the analysis of dialect variation for over three decades [3, 4, 5], but they have invariably functioned by first probing the linguistic differences between each pair of a range of varieties (sites, such as Whitby and Bristol in the UK) over a body of carefully controlled mate-

[1] Wieling and Nerbonne [2] explore whether the perception of dialect differences is subject to a bias toward initial segments in the same way spoken word recognition is, an insight from cognitive science.

rial (say the pronunciation of the vowel in the word 'put'). Second, the techniques AGGREGATE over these linguistic differences, in order, third, to seek the natural groups in the data via clustering or multidimensional scaling (MDS) [6].

Naturally techniques have been developed to determine which linguistic variables weigh most heavily in determining affinity among varieties. But all of the following studies separate the determination of varietal relatedness from the question of its detailed linguistic basis. Kondrak [7] adapted a machine translation technique to determine which sound correspondences occur most regularly. His focus was not on dialectology, but rather on diachronic phonology, where the regular sound correspondences are regarded as strong evidence of historical relatedness. Heeringa [8, pp. 268–270] calculated which words correlated best with the first, second and third dimensions of an MDS analysis of aggregate pronunciation differences. Shackleton [9] used a database of abstract linguistic differences in trying to identify the British sources of American patterns of speech variation. He applied principal component analysis to his database to identify the common components among his variables. Nerbonne [10] examined the distance matrices induced by each of two hundred vowel pronunciations automatically extracted from a large American collection, and subsequently applied factor analysis to the covariance matrices obtained from the collection of vowel distance matrices. Prokić [11] analyzed Bulgarian pronunciation using an edit distance algorithm and then collected commonly aligned sounds. She developed an index to measure how characteristic a given sound correspondence is for a given site.

To study varietal relatedness and its linguistic basis in parallel, we apply bipartite spectral graph partitioning. Dhillon [12] was the first to use spectral graph partitioning on a bipartite graph of documents and words, effectively clustering groups of documents and words simultaneously. Consequently, every document cluster has a direct connection to a word cluster; the document clustering implies

a word clustering and vice versa. In his study, Dhillon [12] also demonstrated that his algorithm identified good document and word clusters.

The usefulness of this approach is not only limited to clustering documents and words simultaneously. For example, Kluger et al. [13] used a somewhat adapted bipartite spectral graph partitioning approach to successfully cluster microarray data simultaneously in clusters of genes and conditions.

There are two main contributions of this paper. The first contribution, which has also been described (in less detail) by Wieling and Nerbonne [1], is to apply a graph-theoretic technique, bipartite spectral graph partitioning, to a new sort of data, namely dialect pronunciation data, in order to solve an important problem, namely how to recognize groups of varieties in this sort of data while simultaneously characterizing the linguistic basis of the group. The second contribution is the application of a ranking procedure to determine the most important sound correspondences with respect to a reference variety in a cluster of varieties. This approach is an improvement over the procedure of ranking the most important elements in a cluster based only on their frequency [12], because it also takes differences between clusters into account.

The remainder of the paper is structured as follows. Section 2 presents the material we studied, a large database of contemporary Dutch pronunciations. Section 3 presents the methods, including the alignment technique used to obtain sound correspondences, the bipartite spectral graph partitioning we used to simultaneously seek affinities in varieties as well as affinities in sound correspondences, and the method to rank the importance of the sound correspondences in each cluster. Section 4 presents our results, while Section 5 concludes with a discussion and some ideas on avenues for future research.

2. Material

In this study we use the most recent broad-coverage Dutch dialect data source available: data from the Goeman-Taeldeman-Van Reenen-project (GTRP) [14, 15]. The GTRP consists of digital transcriptions for 613 dialect varieties in the Netherlands (424 varieties) and Belgium (189 varieties), gathered during the period 1980–1995. For every variety, a maximum of 1876 items was narrowly transcribed according to the International Phonetic Alphabet. The items consist of separate words and phrases, including pronominals, adjectives and nouns. A detailed overview of the data collection is given by Taeldeman and Verleyen [16].

Because the GTRP was compiled with a view to documenting both phonological and morphological variation [17] and our purpose here is the analysis of phonology (pronunciation), we ignore many items of the GTRP. We use the same 562 item subset as introduced and discussed in depth by Wieling et al. [18]. In short, the 1876 item word list was filtered by selecting only single word items, including plural nouns (the singular form was sometimes preceded by an article and therefore not included), base forms of adjectives instead of comparative forms and the first-person plural verb instead of other forms. We omit words whose variation is primarily morphological as we wish to focus on pronunciation. In all varieties the same lexeme was used for a single item.

Because the GTRP transcriptions of Belgian varieties are fundamentally different from transciptions of the Netherlandic varieties as they did not use the same number of phonetic tokens [18], we will restrict our analysis to the 424 Netherlandic varieties. The geographic distribution of these varieties including province names is shown in Figure 1. Furthermore, note that we will not look at diacritics, but only at the 82 distinct phonetic symbols. The average length of every item in the GTRP (without diacritics) is 4.7 tokens (symbols in a phonetic transcription).

Figure 1: Distribution of GTRP localities including province names

3. Methods

To obtain a clear signal of varietal differences in phonology, we ideally want to compare the pronunciations of each variety with a single reference point. We might have used the pronunciations of a proto-language for this purpose, but these are not available in the same transcription system. We settled on using the sound correspondences of a given variety with respect to a reference point as a means of comparison. These sound correspondences form a general and elaborate basis of comparison for the varieties. The use of the correspondences as a basis of comparison is general in the sense that we can determine the correspondences for each variety (more on how we do this below), and it is elaborate since it results in nearly 1000 points of comparison (sound correspondences).

But this strategy also leads to the question of what to use as a reference point. There are no pronunciations in standard Dutch in the GTRP and transcribing the standard Dutch pronunciations ourselves would likely have introduced between-transcriber inconsistencies. Heeringa [8, pp. 274–276] identified pronunciations in the variety of Haarlem as being the closest to standard Dutch. Because Haarlem was not included in the GTRP varieties, we chose the transcriptions of Delft (also close to standard Dutch) as our reference transcriptions. See the discussion section for a consideration of alternatives.

3.1. Obtaining sound correspondences

To obtain the sound correspondences for every site in the GTRP with respect to the reference site Delft, we used an adapted version of the regular Levenshtein algorithm [19].

The Levenshtein algorithm aligns two (phonetic) strings by minimizing the number of edit operations (i.e. insertions, deletions and substitutions) required to

transform one string into the other. For example, the Levenshtein distance between [bɪndən] and [bɛində], two Dutch variants of the word 'to bind', is 3:

bɪndən	insert ɛ	1
bɛɪndən	subst. i/ɪ	1
bɛindən	delete n	1
bɛində		
		3

The corresponding alignment is:

b		ɪ	n	d	ə	n
b	ɛ	i	n	d	ə	
	1	1				1

When all edit operations have the same cost, multiple alignments yield a Levenshtein distance of 3 (i.e. by aligning the [ɪ] with the [ɛ]). To obtain only the best alignments we used an adaptation of the Levenshtein algorithm which uses automatically generated segment substitution costs based on pointwise mutual information (PMI) [20]. This adaptation was proposed, described in detail, and evaluated by Wieling et al. [21] and resulted in significantly better individual alignments than using the regular Levenshtein algorithm.

The approach consists of obtaining initial string alignments by using the Levenshtein algorithm with a syllabicity constraint: vowels may only align with (semi-) vowels, and consonants only with consonants, except for syllabic consonants which may also be aligned with vowels. After the initial run, the substitution cost of every segment pair (a segment can also be a gap, representing insertion and deletion) is calculated according to a pointwise mutual information procedure assessing the statistical dependence between the two segments:

$$\text{PMI}(x, y) = \log_2 \left(\frac{p(x, y)}{p(x)\, p(y)} \right)$$

Where:

- $p(x, y)$ is estimated by calculating the number of times x and y occur at the same position in two aligned strings X and Y, divided by the total number of aligned segments (i.e. the relative occurrence of the aligned segments x and y in the whole data set). Note that either x or y can be a gap in the case of insertion or deletion.

- $p(x)$ and $p(y)$ are estimated as the number of times x (or y) occurs, divided by the total number of segment occurrences (i.e. the relative occurrence of x or y in the whole data set). Dividing by this term normalizes the co-occurrence frequency with respect to the frequency expected if x and y are statistically independent.

Positive PMI values indicate that segments tend to cooccur in correspondences (the greater the PMI value, the more segments tend to cooccur), while negative PMI values indicate that segments do not tend to cooccur in correspondences. New segment distances (i.e. segment substitution costs) are generated by subtracting the PMI value from 0 and adding the maximum PMI value (to ensure that the minimum distance is 0).

After the new segment substitution costs have been calculated, the strings are aligned again based on these new segment substitution costs. Calculating new segment distances and realigning the strings is repeated until the string alignments remain constant. Our alignments were stable after 12 iterations.

After obtaining the final string alignments, we use a matrix to store the presence or absence of each segment substitution for every variety (with respect to the

reference place). We thus obtain a binary $m \times n$ matrix $\boldsymbol{A}$ of m varieties (in our case 423; Delft was excluded as it was used as our reference site) by n segment substitutions (in our case 957; not all possible segment substitutions occur). A value of 1 in $\boldsymbol{A}$ (i.e. $A_{ij} = 1$) indicates the presence of segment substitution j in variety i (compared to the reference variety), while a value of 0 indicates the absence. We experimented with frequency thresholds, but decided against applying one in this paper as their application seemed to lead to poorer results. We postpone a consideration of frequency-sensitive alternatives to the discussion section.

3.2. Bipartite spectral graph partitioning

An undirected bipartite graph can be represented by $G = (R, S, E)$, where R and S are two sets of vertices and E is the set of edges connecting vertices from R to S. There are no edges between vertices in a single set, e.g. connecting nodes in R. In our case R is the set of varieties, while S is the set of sound segment substitutions (i.e. sound correspondences). An edge between r_i and s_j indicates that the sound segment substitution s_j occurs in variety r_i. It is straightforward to see that matrix $\boldsymbol{A}$ is a representation of an undirected bipartite graph. Figure 2 shows an example of an undirected bipartite graph consisting of four varieties and three sound correspondences.

If we represent a graph such as that in Figure 2 using a binary adjacency matrix in which a cell (a,b) has the value 1 just in case there is an edge from a to b, and 0 otherwise, then the spectrum of the graph is the set of eigenvalues of its adjacency matrix. Note that the adjacency matrix (having $(m + n) \times (m + n)$ elements) is larger than $\boldsymbol{A}$ (having $m \times n$ elements), as it contains values for all possible vertex combinations.

Spectral graph theory is used to find the principal properties and structure of a graph from its graph spectrum [22]. Dhillon [12] was the first to use spectral graph

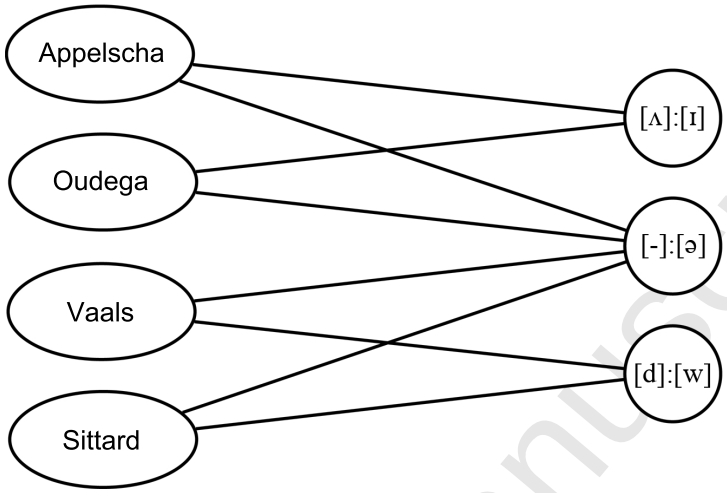


Figure 2: Example of a bipartite graph of four varieties and three sound correspondences

partitioning on a bipartite graph of documents and words, effectively clustering groups of documents and words simultaneously. Consequently, every document cluster has a direct connection to a word cluster. In similar fashion, we would like to obtain a clustering of varieties and corresponding segment substitutions.

The algorithm of Dhillon [12] is based on the fact that finding the optimal bipartition having balanced clusters is solved by finding the eigenvector corresponding with the second smallest eigenvalue of the adjacency matrix. Using linear algebra, it turns out that this solution can also be found by computing the left and right singular vectors corresponding to the second (largest) singular value of the normalized word-by-document matrix (or in our case variety-by-segment-correspondence matrix). Because the latter matrix is smaller than the adjacency matrix needed for the first method, the second method is computationally much cheaper.

Instead of finding only two clusters, it is also possible to find k clusters, using $l = \lceil \log_2 k \rceil$ singular vectors. The steps needed for this approach are illustrated below in the multipartioning algorithm introduced and explained in more detail by

Dhillon [12].

1. Given the $m \times n$ variety-by-segment-correspondence matrix $\boldsymbol{A}$ as discussed previously, form the normalized matrix

$$\boldsymbol{A_n} = \boldsymbol{D_1}^{-1/2} \boldsymbol{A} \boldsymbol{D_2}^{-1/2}$$

with $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$ diagonal matrices such that $D_1(i,i) = \Sigma_j A_{ij}$ and $D_2(j,j) = \Sigma_i A_{ij}$

2. Calculate the singular value decomposition (*SVD*) of the normalized matrix $\boldsymbol{A_n}$ to obtain the singular values ($\boldsymbol{\Lambda}$) and the left (the columns of $\boldsymbol{U}$) and right (the columns of $\boldsymbol{V}$) singular vectors

$$\textit{SVD}(\boldsymbol{A_n}) = \boldsymbol{U} * \boldsymbol{\Lambda} * \boldsymbol{V}^{\mathrm{T}}$$

3. Calculate the $l = \lceil \log_2 k \rceil$ eigenvectors needed for the partitioning based on the singular vectors

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{D_1}^{-1/2} \boldsymbol{U}_{[\mathbf{2},\ldots,l+\mathbf{1}]} \\ \boldsymbol{D_2}^{-1/2} \boldsymbol{V}_{[\mathbf{2},\ldots,l+\mathbf{1}]} \end{bmatrix}$$

4. Run the k-means algorithm on $\boldsymbol{Z}$ to obtain the k-way multipartitioning

To illustrate this procedure, we will co-cluster the following variety-by-segment-correspondence matrix $\boldsymbol{A}$ in $k = 2$ groups (note that this matrix is visualized by Figure 2).

	[ʌ]:[ɪ]	[-]:[ə]	[d]:[w]
Appelscha (Friesland)	1	1	0
Oudega (Friesland)	1	1	0
Vaals (Limburg)	0	1	1
Sittard (Limburg)	0	1	1

We first construct matrices $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$. $\boldsymbol{D_1}$ contains the total number of edges from every variety (in the same row) on the diagonal, while $\boldsymbol{D_2}$ contains the total number of edges from every segment substitution (in the same column) on the diagonal. Both matrices are shown below.

$$\boldsymbol{D_1} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \qquad \boldsymbol{D_2} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

We can now calculate $\boldsymbol{A_n}$ using the formula displayed in step 1 of the multipartitioning algorithm:

$$\boldsymbol{A_n} = \begin{bmatrix} .5 & .35 & 0 \\ .5 & .35 & 0 \\ 0 & .35 & .5 \\ 0 & .35 & .5 \end{bmatrix}$$

Applying the *SVD* to $\boldsymbol{A_n}$ yields:

$$\boldsymbol{U} = \begin{bmatrix} -.5 & .5 & .71 & 0 \\ -.5 & .5 & -.71 & 0 \\ -.5 & -.5 & 0 & -.71 \\ -.5 & -.5 & 0 & .71 \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & .71 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{V}^{\mathrm{T}} = \begin{bmatrix} -.5 & -.71 & -.5 \\ .71 & 0 & -.71 \\ -.5 & .71 & -.5 \end{bmatrix}$$

To cluster in two groups, we look at the second singular vector of $\boldsymbol{U}$ (second column) and $\boldsymbol{V}^{\mathrm{T}}$ (second row; i.e. second column of $\boldsymbol{V}$) and compute the 1-dimensional vector $\boldsymbol{Z}$:

$$\boldsymbol{Z} = \begin{bmatrix} .35 & .35 & -.35 & -.35 & .5 & 0 & -.5 \end{bmatrix}^T$$

Note that the first four values correspond with the places (Appelscha, Oudega, Vaals and Sittard) and the final three values correspond to the segment substitutions ([ʌ]:[ɪ], [-]:[ə] and [d]:[w]).

After running the k-means algorithm (with random initialization) on $\boldsymbol{Z}$, where $k = 2$, the items are assigned to one of two clusters as follows:

$$\begin{bmatrix} 1 & 1 & 2 & 2 & 1 & 1 & 2 \end{bmatrix}^T$$

The clustering shows that Appelscha and Oudega are clustered together (corresponding to the first and second item of the vector, above) and linked to the clustered segment substitutions of [ʌ]:[ɪ] and [-]:[ə] (cluster 1). Similarly, Vaals and Sittard are clustered together and linked to the clustered segment substitution [d]:[w] (cluster 2). Note that the segment substitution [-]:[ə] (an insertion of [ə]) is actually not meaningful for the clustering of the varieties (as can also be observed in $\boldsymbol{A}$), because the middle value of $\boldsymbol{V}^{\mathrm{T}}$ corresponding to this segment substitution is 0. It could therefore just as likely be grouped in cluster 2. Nevertheless, the k-means algorithm always assigns every item to a single cluster.

The procedure to determine the importance of sound correspondences in a cluster is discussed next.

3.3. Determining the importance of sound correspondences

Before deciding how to calculate the importance of each sound correspondence, we need to consider the characteristics of important sound correspondences.

Note that if a variety contains a sound correspondence, this simply means that the sound correspondence (i.e. two aligned segments) occurs at least once in any of the aligned pronunciations (with respect to the reference variety Delft).

In the following, we will discuss two characteristics of an important sound correspondence, 'representativeness' and 'distinctiveness'.

Representativeness indicates the proportion of varieties in the cluster which contain the sound correspondence. A value of 0 indicates that the sound correspondence does not occur in any of the varieties, while a value of 1 indicates that the sound correspondence occurs in all varieties in the cluster. This is shown in the formula below for sound correspondence [a]:[b] and cluster c_i:

$$\text{Representativeness}(\text{a}, \text{b}, c_1) = \frac{\text{number of varieties in } c_i \text{ containing [a]:[b]}}{\text{total number of varieties in } c_i}$$

The second characteristic of an important sound correspondence is 'distinctiveness'. This characteristic indicates how prevalent a sound correspondence is in its own cluster as opposed to other clusters.

Suppose sound correspondence [a]:[b] is clustered in group c_1. We can count how many varieties in c_1 contain sound correspondence [a]:[b] and how many varieties in total contain [a]:[b]. Dividing these two values yields the relative occurrences of [a]:[b] in c_1.

$$\text{RelativeOccurrence}(\text{a}, \text{b}, c_1) = \frac{\text{number of varieties in } c_1 \text{ containing [a]:[b]}}{\text{number of varieties containing [a]:[b]}}$$

For instance, if [a]:[b] occurs in 20 varieties of which 18 belong to c_1, the relative occurrence is 0.9. We intend to capture in this measure how well the correspondence signals the area represented by c_1. While it may seem that this number can tell us if a sound correspondence is distinctive or not, this is not the case. For instance, if c_1 consists of 95% of all varieties, the sound correspondence [a]:[b] is not very distinctive for c_1 (i.e. we would expect [a]:[b] to occur in 19 varieties instead of 18). To correct for this, we also need to take into account the relative size of c_1.

$$\text{RelativeSize}(c_1) = \frac{\text{number of varieties in } c_1}{\text{total number of varieties}}$$

We can now calculate the distinctiveness of a sound correspondence by subtracting the relative size from the relative occurrence. Using the previous example, this would yield $0.90 - 0.95 = -0.05$. A positive value indicates that the sound correspondence is distinctive (the higher the value, the more distinctive), while a negative value indicates values which are not distinctive. To ensure the maximum value equals 1, we use a normalizing term as can be seen in the following formula:

$$\text{Distinctiveness}(\text{a}, \text{b}, c_1) = \frac{\text{RelativeOccurrence}(\text{a}, \text{b}, c_1) - \text{RelativeSize}(c_1)}{1 - \text{RelativeSize}(c_1)}$$

A distinctiveness value of 0 indicates that the observed and expected percentage are equal. Values below 0 (which are unbounded) indicate sound correspondences which are not distinctive, while positive values indicate distinctive values.

To be able to rank the sound correspondences based on their distinctiveness and representativeness, we need to combine these two values. A simple way to deter-

mine the importance of every sound correspondence based on the distinctiveness and representativeness is to take the average of both values, as is illustrated in the following formula.

$$\text{Importance}(\text{a}, \text{b}, c_1) = \frac{\text{Representativeness}(\text{a}, \text{b}, c_1) + \text{Distinctiveness}(\text{a}, \text{b}, c_1)}{2}$$

It is clear that we might explore more complicate combinations, but we regard both representativeness and distinctiveness as equally important.

Because it is essential for an important sound correspondence to be distinctive, we will only consider sound correspondences having a non-negative distinctiveness. As both representativeness and distinctiveness will now range between 0 and 1, the importance will also range between 0 and 1. Higher values *within a cluster* indicate more important sound correspondences for that cluster. Since we took the cluster size into account in calculating the distinctiveness, we can also compare the clusters with respect to the importance values of their sound correspondences.

In the following section we will report the results on clustering in two, three and four groups.[2]

[2]We also experimented with clustering in more than four groups, but the k-means clustering algorithm did not give stable results for these groupings. It is possible that the random initialization of the k-means algorithm caused the instability of the groupings, but since we are ignoring most of the information contained in the alignments it is more likely that this causes a decrease in the number of clusters we can reliably detect.

4. Results

After running the multipartitioning algorithm[3] on the variety-by-segment correspondence matrix we obtained a two-way clustering in k clusters of varieties and segment substitutions. Figure 3 illustrates the simultaneous clustering in two dimensions. A black dot is drawn if the variety (y-axis) contains the segment substitution (x-axis). The varieties and segments are sorted in such a way that the clusters are clearly visible (and marked) on both axes.

To visualize the clustering of the varieties, we created geographical maps in which we indicate the cluster of each variety by a distinct pattern. The division in 2, 3 and 4 clusters is shown in Figure 4. We note that the four-way split in Figure 4 does not respect the divisions inferred in the three-way split. Our work inherits this property from the k-means clustering algorithm, which always seeks the optimal division into k groups. It is clear that this divistion does not need to respect the groupings proposed when creating $k - 1$ groups.

In the following subsections we will discuss the most important geographical clusters together with their simultaneously derived sound correspondences. In the earlier paper [1] we discussed several sound correspondences which we recognized from the handbooks on Dutch dialectology, such as [23]. We discuss these examples here as well, but we go on to suggest how to determine the most important correspondences quantitatively, and inspect the results of applying our suggestions to the Dutch data. The main point to note is that besides a sensible geographical clustering, we also obtain linguistically sensible results.

Note that the connection between a cluster of varieties and sound correspondences does not necessarily imply that those sound correspondences occur only in

[3]The implementation of the multipartitioning algorithm was obtained from `http://adios.tau.ac.il/SpectralCoClustering`

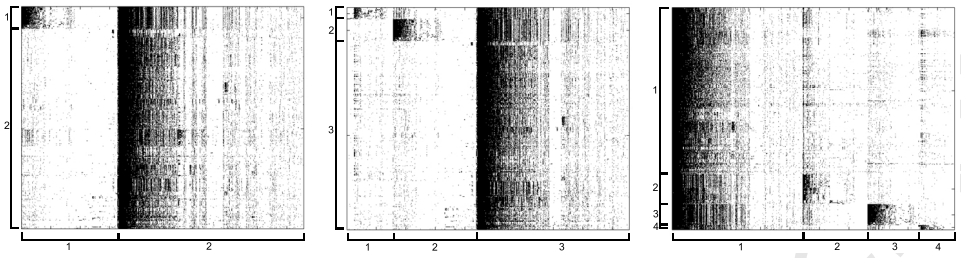


Figure 3: Co-clustering varieties (y-axis) and segment substitutions (x-axis) in 2 (left), 3 (middle) and 4 (right) clusters

that particular cluster of varieties. This can also be observed in Figure 3, where sound correspondences in a particular cluster of varieties also appear in other clusters (but less densely).

The Frisian area

The division into two clusters clearly separates the Frisian language area (in the province of Friesland) from the Dutch language area. This is the expected result as Heeringa [8, pp. 228–229] also measured Frisian as the most distant of all the language varieties spoken in the Netherlands and Flanders. It is also expected in light of the fact that Frisian even has the legal status of a different language rather than a dialect of Dutch. Note that the separate "islands" in the Frisian language area (see Figure 4) correspond to the Frisian cities which are generally found to deviate from the rest of the Frisian language area [8, pp. 235–241].

A few interesting sound correspondences between the reference variety (Delft) and the Frisian area are displayed in the following table and discussed below. The correspondences in the table were chosen subjectively from the long list of correspondences provided by the bipartite spectral graph clustering procedure based on their being interpretable, e.g. based on the literature concerning Dutch and Frisian [1]. We also compare the correspondences selectively chosen in earlier work with

Figure 4: Geographic projection of varieties clustered in 2 clusters (left), 3 clusters (middle) and 4 clusters (right)

those determined by the calculations based on importance, i.e. the combination of representativity and distinctiveness. The importance of each subjectively selected sound correspondence is indicated in the table by its rank (the most important sound correspondence has rank 1). For completeness, the importance and its two components distinctiveness and representativeness are also displayed.

Reference	[ʌ]	[ʌ]	[a]	[o]	[u]	[x]	[x]	[r]
Frisian	[ɪ]	[i]	[i]	[ɛ]	[ɛ]	[j]	[z]	[x]
Rank	**3**	**8**	20	**1**	11	14	**4**	41
Importance	0.96	0.88	0.77	0.97	0.85	0.83	0.96	0.65
Representativeness	0.95	0.76	1	1	0.86	1	0.95	0.81
Distinctiveness	0.97	1	0.54	0.95	0.84	0.65	0.97	0.49

Only looking at the sound correspondences *pur sang*, we can see that the Dutch /a/ or /ʌ/ is pronounced [i] or [ɪ] in the Frisian area. This well known sound correspondence can be found in words such as *kamers* 'rooms', Frisian [kɪməs] (pro-

nunciation from Anjum), or *draden* 'threads' and Frisian [trɪdn] (Bakkeveen). In addition, the Dutch (long) /o/ and /u/ both tend to be realized as [ɛ] in words such as *bomen* 'trees', Frisian [bjɛmən] (Bakkeveen) or *koeien* 'cows', Frisian [kɛi] (Appelscha).

We also identify clustered correspondences of [x]:[j] where Dutch /x/ has been lenited, e.g. in *geld* (/xɛlt/) 'money', Frisian [jɪlt] (Grouw), but note that [x]:[g] as in [gɛlt] (Sint-Annaparochie) also occurs, illustrating that sound correspondences from another cluster (i.e. the rest of the Netherlands) can indeed also occur in the Frisian area. Another sound correspondence co-clustered with the Frisian area is the Dutch /x/ and Frisian [z] in *zeggen* (/zɛxə/) 'say' Frisian [sizə] (Appelscha).

Besides the previous results, we also note some problems. First, the accusative first-person plural pronoun *ons* 'us' lacks the nasal in Frisian, but the correspondence was not tallied in this case because the nasal consonant is also missing in Delft. Second, some apparently frequent sound correspondences result from historical accidents, e.g. [r]:[x] corresponds regularly in the Dutch:Frisian pair [dor]:[trux] 'through'. Dutch has lost the final [x] and either Dutch or Frisian has experienced metathesis in the vowel-[r] sequence. These two sorts of examples might be treated more satisfactorily if we were to compare pronunciations not to a standard language, but rather to a reconstruction of a proto-language.

When looking at the importance of the subjectively chosen sound correspondences, we can see that out of the eight selected sound correspondences four are in the top ten. As there are a total of 184 sound correspondences grouped in the Frisian cluster, this indicates that our method to identify important sound correspondences conforms largely with our subjectively selected important sound correspondences. In addition, the problematic sound correspondence [r]:[x] has a very low rank. Also note that the sound correspondence [a]:[i] has a low rank (20), but clearly belongs to a group of important sound correspondences consisting of /a/-

and /i/-like sounds (i.e. [ʌ]:[ɪ] has rank 3 and [ʌ]:[i] has rank 8).

The table below shows the 5 most important sound correspondences for the Frisian area ranked according to their importance.

Rank	1	2	3	4	5
Importance	0.97	0.96	0.96	0.96	0.94
Reference	[o]	[o]	[ʌ]	[x]	[k]
Frisian	[ɛ]	[ɪ]	[ɪ]	[z]	[ʃ]

Besides the sound correspondences already discussed, we see two additional important sound correspondences, [o]:[ɪ] and [k]:[ʃ]. The former is illustrated in the word *schoon* [sxo:n] 'clean', pronounced [skɪn] (Bakkeveen) or [skjɪn] (Holwerd, Joure), as well as in *kopen* [ko:pə] 'buy', pronounced [kɪəpjə] (Grouw, Akkrum). The latter correspondence occurs in words such as *raken* [ra:kə] 'strike' pronounced [rɛitʃə] in Friesland (Appelscha, Bakkeveen) or in *maken* [ma:kə] 'make' pronounced [mɛitʃə] (e.g. in Anjum and Jubbega). The first may be part of the larger shift noted above, in which the low back vowels have been fronted, and the latter is the result of a palatalization of the /k/, which is commonly noted in Frisian.

The Limburg area

The division into three clusters separates the southern Limburg area from the rest of the Dutch and Frisian language area. This result is also in line with previous studies investigating Dutch dialectology; Heeringa [8, pp. 228–229] found the Limburg dialects to deviate most strongly from other different dialects within the Netherlands-Flanders language area once Frisian was removed from consideration.

Some interesting sound segment correspondences for Limburg are displayed in the following table and discussed below. See [24] for further discussion of sound correspondences involving Limburg.

Reference	[r]	[r]	[k]	[n]	[n]	[w]
Limburg	[ʀ]	[ʁ]	[x]	[ʀ]	[ʁ]	[f]
Rank	12	21	**1**	**5**	13	57
Importance	0.58	0.53	0.81	0.64	0.57	0.44
Representativeness	1	0.91	0.86	0.77	0.59	0.68
Distinctiveness	0.16	0.15	0.75	0.51	0.54	0.21

Looking only at the sound correspondences, Limburg uses more uvular versions of /r/, i.e. the trill [ʀ], but also the voiced uvular fricative [ʁ]. These occur in words such as *over* 'over, about', but also in *breken* 'to break', i.e. both pre- and post-vocalically. The bipartite clustering likewise detected examples of the famous "second sound shift", in which Dutch /k/ is lenited to /x/, e.g. in *ook* 'also' realized as [ox] in Epen and elsewhere. Interestingly, when looking at other words there is less evidence of lenition in the words *maken* 'to make', *gebruiken* 'to use', *koken* 'to cook', and *kraken* 'to crack', where only two Limburg varieties document a [x] pronunciation of the expected stem-final [k], namely Kerkrade and Vaals. The limited linguistic application does appear to be geographically consistent, but Kerkrade pronounces /k/ as [x] where Vaals lenites further to [s] in words such as *ruiken* 'to smell', *breken* 'to break', and *steken* 'to sting'. Further, there is no evidence of lenition in words such as *vloeken* 'to curse', *spreken* 'to speak', and *zoeken* 'to seek', which are lenited in German (*fluchen, sprechen, suchen*).

Some regular correspondences merely reflect other, and sometimes more fundamental differences. For instance, we found correspondences between [n] and [ʀ] or [ʁ] for Limburg , but this turned out to be a reflection of the older plurals in -r. For example, in the word *wijf* 'woman', plural *wijven* in Dutch, *wijver* in Limburg dialect. Dutch /w/ is often realized as [f] in the word *tarwe* 'wheat', but this is also due to the elision of the final schwa, which results in a pronunciation such as

[taʀəf], in which the standard final devoicing rule of Dutch is applicable.

While there was much overlap between the subjectively selected and the best objectively determined sound correspondences in the Frisian area, this appears to be less so for the sound correspondences of Limburg. Out of the top ten sound correspondences, we only identified two sound correspondences. However, considering that there are 156 sound correspondences in the Limburg cluster, we can still conclude that we detected many good sound correspondences (four out of six are still in the top fifteen). Just as in the Frisian cluster, we can also identify a group of sound correspondences here, linking an /r/-like sound in Delft to a slightly different /r/-like sound in Limburg (i.e. [r]:[ʀ] and [r]:[ʁ]).

The table below shows the 5 most important sound correspondences for the Limburg area ranked according to their importance. When comparing this table to the top five sound correspondences of Friesland, we clearly see that Friesland is characterized by more important sound correspondences than Limburg ($t = 8.4$, $p < .001$). This is a sensible result as in Friesland a separate language is spoken, while in the greater part of Limburg only a dialect of Dutch is spoken.

Rank	1	2	3	4	5
Importance	0.81	0.73	0.67	0.65	0.64
Reference	[k]	[n]	-	[o]	[n]
Limburg	[x]	[x]	[p]	[y]	[ʀ]

Besides the sound correspondences discussed above, we see three additional important sound correspondence, [n]:[x], -:[p] (an insertion of a [p]) and [o]:[y]. The [n]:[x] correspondence is ultimately morphological, we suspect, but it appears in words such as *bladen* [bla:dən] ‘magazines’ pronounced [blajəx] in e.g. Roermond and Venlo in Limburg; *graven* [xra:vən] ‘graves’, pronounced [xʁavəx] in

Gulpen, Horn and elsewhere in Limburg; and *kleden* [kledən] '(table)cloths', pronounced [klɛɪdəx] e.g. in Kerkrade and Gulpen in Limburg. The [p] is interpreted as inserted in the words *krom* 'crooked' Limburg [kʁɔmp] (Echt) and *komen* 'come [inf.]' Limburg [køɔmp] (Vaals). The correspondence [o]:[y] is found in words such as *droog* 'dry' Limburg [dʁyx] (Eisden); *bomen* 'trees' Limburg [byəm] (Sevenum); *geloven* 'believe [3rd.pl.pres.]' Limburg [xəlyəvə] (Vaals); *dopen* 'baptize' Limburg [dyəpə] (Meijel); and *dooien* 'thaw' Limburg [tyənə] (Kerkrade).

The Low Saxon area

Finally, the division into four clusters also separates the varieties from Groningen and Drenthe from the rest of the Netherlands. This result differs somewhat from the standard scholarship on Dutch dialectology [8], according to which the Low Saxon area should include not only the provinces of Groningen and Drenthe, but also the province of Overijssel and the northern part of the province of Gelderland. We do not know the reason for this difference, but it might be that the simplifying steps needed for our method (i.e. we only consider the presence or absence of sound correspondences) are of influence.

A few interesting sound correspondences [23] are displayed in the following table and discussed below.

Reference	[ə]	[ə]	[ə]	-	[a]
Low Saxon	[m]	[ŋ]	[ɴ]	[ʔ]	[e]
Rank	**7**	**8**	58	53	14
Importance	0.59	0.58	0.43	0.51	0.56
Representativeness	0.98	0.98	0.12	0.88	0.88
Distinctiveness	0.20	0.18	0.74	0.13	0.25

Looking only at the sound correspondences, the best known characteristic of this area, the so-called "final n" (*slot n*) is instantiated strongly in words such as *strepen*, 'stripes', realized as [strepm] in the northern Low Saxon area. It would be pronounced [strepə] in standard Dutch, so the differences shows up as an unexpected correspondence of [ə] with [m], [ŋ] and [ɴ]. We return to this below.

The pronunciation of this area is also distinctive in normally pronouncing words with initial glottal stops [ʔ] rather than initial vowels, e.g. *af* 'finished' is realized as [ʔɔf] (Schoonebeek). Furthermore, the long /a/ is often pronounced in an umlauted fashion, [e], as in *kaas* 'cheese', [kes], e.g. in Gasselte, Hooghalen and Norg.

In contrast to the other two clusters, we did not subjectively select any top five sound correspondences. We did, however, select two top ten sound correspondences out of a total of 220 sound correspondences in the Low Saxon cluster. This again indicates that our importance ranking method overlaps somewhat with our subjective selection. Similar to the other two clusters, we can identify a group of sound correspondences, consisting of the /ə/-like sound in Delft connected to an /m/- or /n/-like sound in Low Saxon (i.e. [ə]:[m], [ə]:[ŋ] and [ə]:[ɴ], mentioned above).

The table below shows the 5 most important sound correspondences for the Low Saxon area ranked according to their importance. It is clear that the Low Saxon area is characterized by sound correspondences which are somewhat less important than the sound correspondences for Limburg ($t = 3.5$, $p < .002$) and much less important than the sound correspondences in Friesland ($t = 12.4$, $p < .001$). This again is a sensible result as the dialect spoken in the Low Saxon area is less pronounced than the dialect of Limburg.

Rank	1	2	3	4	5
Importance	0.73	0.68	0.63	0.60	0.60
Reference	[f]	[v]	[f]	[k]	[ʋ]
Low Saxon	[b]	[b]	[m]	[ʔ]	[b]

Unfortunately we did not detect any top five sound correspondences, hence all top five sound correspondences will be discussed next. The [f]:[b] correspondence appears in words such as *proeven* [prufə] 'test', pronounced [proybm̩] in Low Saxon, in e.g. Barger-Oosterverld and Bellingwolde; *schrijven* [sxrəɪfə] pronounced [sxribm̩] in Low Saxon, e.g. in Aanloo and Aduard; or *wrijven* [frəɪfə] 'rub', pronounced [vribm̩] in Low Saxon. Similar examples involve *schuiven* 'shove' and *schaven* 'scrape, plane'. Note that the diphtong in [sxrəɪfə] and [frəɪfə] is not standard Dutch, but rather the pronunciation in our reference variety Delft.

The correspondence just discussed involves the lenition and devoicing of the stop consonant /b/, but we also have examples where the /b/ has been lenited, but not devoiced, and this is the second correspondence, to which we now turn. In these cases [v] corresponds with the [b] in words such as *leven* 'live [3rd.pl.pres.]' Low Saxon [lebm̩] (Aduard); *bleven* 'remain [3rd.pl]' Low Saxon [blibm̩] (Aduard); *doven* 'deaf [pl.]' Low Saxon [dubm̩] (Aduard); *graven* 'dig [3rd.pl.]' Low Saxon [xrabm̩] (Anloo); and *dieven* 'thieves' Low Saxon [dibm̩] (Westerbork). In all these cases we encounter a [v] in the reference variety in place of the [b]. The [f]:[m] correspondence occurs in words such as *zeventig* 'seventy' Low Saxon [zømtix] (Aduard); *boven* 'above' Low Saxon [bom] (Aduard); *proeven* 'taste, sample [3rd.pl.pres]' Low Saxon [prym] (Hollandsche Veld); *schrijven* 'write [3rd. pl.pres]' Low Saxon [sxrim] (Bellingwolde); and *wrijven* 'rub [3rd.pl.pres]' Low Saxon [ʋrim] (Aduard).

The final two examples are interesting because they are not commented on

often. We notice [k]:[ʔ] correspondence where /k/ was historically initial in the weak syllable /-kən/, for example in *planken* [plɑŋkə] 'boards, planks', pronounced [plɑŋʔŋ̩] in Low Saxon. Similar examples are provided by *denken* 'think' pronounced [dɛŋʔŋ̩] in Low Saxon, and *drinken* 'drink' pronounced [drɪŋʔŋ̩] in Low Saxon. This correspondence is quite regular, while the last, [ʊ]:[b], is found in exactly one word *eeuwen* [euʊə] 'centuries', pronounced [eubm̩] at several Low Saxon sites.

We note that some of these examples again detect likely non-historical correspondences, e.g. the [f]:[m] correspondence exemplified in the Aduard pronunciation [bom], where the [m] may correspond not to the medial Delft [f] ([bofən]), but rather to the final [n], which has assimilated in place due to the [f], which in turn was then later lost. The same may be said of the examples *proeven* 'taste, sample' and *schrijven* 'write' (above). This underscores the need for manual inspection even though the procedure is successful in uncovering interesting correspondences.

5. Discussion

In this study, we have applied a novel method to dialectology in simultaneously determining groups of varieties and their linguistic basis (i.e. sound segment correspondences). We demonstrated that the bipartite spectral graph partitioning method introduced by Dhillon [12] gave sensible clustering results in the geographical domain as well as for the concomitant linguistic basis.

We are optimistic that the use of techniques such as the one presented in this paper can be more successful in engaging traditional dialectologists exactly because the relation between the proposed division into dialect areas and the linguistic basis of the division is directly accessible. This is not the case in dialectometry work in which aggregate relations form the basis for the division into dialect areas. In

those approaches, linguistic differences are summed and the sum of differences is extracted as an indicator of the relations between varieties. This perspective has its advantages [6], but it has not converted large numbers of dialectologists to the use of exact techniques. While the techniques in the present paper are more complicated, their linguistic basis is more accessible. There are nonetheless several points where improvements to the present approach would be welcome. It is unclear now which of these will be easily incorporated and which will be difficult.

As mentioned above, we did not have transcriptions of standard Dutch, but instead we used transcriptions of a variety (Delft) close to the standard language. While the pronunciations of most items in Delft were similar to standard Dutch, there were also items which were pronounced differently from the standard. While we do not believe that this will change our results significantly, using standard Dutch transcriptions produced by the transcribers of the GTRP corpus would make the interpretation of sound correspondences more straightforward.

We indicated in Section 4 that some sound correspondences, e.g. [r]:[x], would probably not occur if we used a reconstructed proto-language as a reference instead of the standard language. A possible way to reconstruct such a proto-language is by multiple aligning [25] all pronunciations of a single word and use the oldest sound at each position [26] in the reconstructed word. Another option is to convert words from a Proto-Germanic dictionary [27] to the same transcription system. It would be interesting to see if using such a reconstructed proto-language would improve the results by removing sound correspondences such as [r]:[x].

We did not try to use another variety as our reference point, as we think our decision to choose a variety close to the standard is the most sensible considering the available data. However, it would be interesting to investigate the effect of changing the reference variety to a variety within one of the three identified clusters (i.e. Friesland, Limburg or Low Saxon). It is conceivable that this will re-

sult in different clusters, since a different set of extracted sound correspondences is used. For example, when choosing a variety in Friesland as our reference point, the sound correspondence [o]:[y] (which is characteristic for Limburg) will probably occur less often, since Frisian varieties frequently use [ɛ] or [ɪ] instead of [o] (see Section 4).

In this study, we only looked at the importance of individual sound correspondences, and we have compared sound correspondences noted in earlier manual work with sound correspondences ranked by simple measures we developed to find representative and distinctive elements. Earlier we had detected important correspondences by examining the clustered sound correspondences and then "eyeballing" sets of alignments to see where they were instantiated. In this paper we have implemented statistical measures designed to find important sound correspondences more automatically. When we then applied these measures to the same data from which we had manually identified what seemed like important correspondences, i.e. the three clusters in Section 4, we note that in every case additional correspondences were noted, which could in turn be confirmed in a manual verification step. We conclude from this that the measures are functioning well. It also turns out that the correspondences noted earlier in the "eyeballing" fashion were also ranked highly, although not always in the top five.

The important sound correspondences found by our procedure are not always the historical correspondences which diachronic linguists build reconstructions on. Instead, they may reflect entire series of sound changes and may involve elements that do not correspond historically at all. We suspect that dialect speakers likewise fail to perceive such correspondences as *general* indicators of another speaker's provenance, except in the specific context of the words such as those in the data set from which the correspondences are drawn. This means that some manual investigation is still necessary to analyze the distinctive elements of the dialects as

well.

We nonetheless note that we may be able to improve the procedures applied here incorporating more phonological context into the basic alignment routines. We have experimented with incorporating more context in the past, and have concluded not only that it is feasible, but also that the result is in general an improvement [28].

While sound segment correspondences function well as a linguistic basis, it might also be fruitful to investigate morphological distinctions present in the GTRP corpus. This would enable us to compare the similarity of the geographic distributions of pronunciation variation on the one hand and morphological variation on the other.

In this study, we focused on the original bipartite spectral graph partitioning algorithm introduced by Dhillon [12]. Investigating other approaches such as biclustering algorithms for biology [29] or an information-theoretic co-clustering approach [30] would be highly interesting.

It would likewise be interesting to attempt to incorporate frequency, by weighting correspondences that occur frequently more heavily than those which occur only infrequently. While it stands to reason that more frequently encountered variation would signal dialectal affinity more strongly, it is also the case that *inverse* frequency weightings have occasionally been applied [4], and have been shown to function well. We have the sense that the last word on this topic has yet to be spoken, and that empirical work would be valuable.

Our paper has improved the techniques available for studying the social dynamics of language variation. In dialect geography, social dynamics are operationalized as geography, and bipartite spectral graph partitioning has proven itself capable of detecting the effects of social contact, i.e. the latent geographic signal in the data. In the future, techniques that attempt not just to detect the geographical signal in

the data, but moreover to incorporate geography as an explict parameter in models of language variation may be in a position to overcome weakness inherent in current models. The work presented here aims only at detecting the geographic signal.

Other dialectometric techniques have done this as well. But linguists have rightly complained that linguistic factors have been neglected in dialectometry [31, p.176]. This paper has shown that bipartite graph clustering can detect the linguistic basis of dialectal affinity, and thus provide the information that Schneider and others have missed.

We additionally note that we may expect that cognitive constraints are also reflected in the empirical basis for studies such as these. They should emerge as strong associations that are *not* conditioned by geography. By showing how to identify geographically conditioned variable associations (correlations), we should be in an improved position to detect the unconditioned ones. One major striking effect in variation due to cognitive dynamics is worth special mention even though we see no way to investigate it using the techniques of the present paper, and that is the fact that sound correspondences are interesting objects of study. This is only possible because sounds are organized phonemically—so that a shift in the pronunciation of a sound in one word tends to result in its shift in another as well. This tendency is real, and is worthy of more exact attention.

Future work will also need to address how cognitive and social dynamics interact, preferably by deploying techniques capable not only of detecting social and cognitive conditioning, but also capable of linking these to the concrete effects they have on the distributions of linguistic features.

Acknowledgments

We thank Assaf Gottlieb for sharing the implementation of the bipartite spectral graph partitioning method. We also would like to thank Peter Kleiweg for supplying the L04 package which was used to generate the maps in this paper. Finally, we are grateful to the anonymous reviewers for their helpful comments and suggestions.

References

[1] M. Wieling, J. Nerbonne, Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology, in: M. Choudhury, S. Hassan, A. Mukherjee, S. Muresan (Eds.), Proc. of the 2009 Workshop on Graph-based Methods for Natural Language Processing, 2009, pp. 26–34.

[2] M. Wieling, J. Nerbonne, Dialect pronunciation comparision and spoken word recognition, in: P. Osenova et al. (Ed.), Proc. of the RANLP Workshop on Computational Phonology Workshop at Recent Advances in Natural Language Processing, RANLP, Borovetz, 2007, pp. 71–78.

[3] J. Séguy, La dialectométrie dans l'atlas linguistique de gascogne, Rev Linguist Roman 37 (145) (1973) 1–24.

[4] H. Goebl, Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie, Österreichische Akademie der Wissenschaften, Wien, 1982.

[5] J. Nerbonne, W. Heeringa, P. Kleiweg, Edit distance and dialect proximity, in: D. Sankoff, J. Kruskal (Eds.), Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed., CSLI, Stanford, CA, 1999, pp. v–xv.

[6] J. Nerbonne, Data-driven dialectology, Lang linguist compass 3 (1) (2009) 175–198.

[7] G. Kondrak, Determining recurrent sound correspondences by inducing translation models, in: Proc. of the Nineteenth International Conference on Computational Linguistics (COLING 2002), COLING, Taipei, 2002, pp. 488–494.

[8] W. Heeringa, Measuring Dialect Pronunciation Differences using Levenshtein Distance, Ph.D. thesis, Rijksuniversiteit Groningen (2004).

[9] R. G. Shackleton, Jr., English-american speech relationships: A quantitative approach, Journal of English Linguistics 33 (2) (2005) 99–160.

[10] J. Nerbonne, Identifying linguistic structure in aggregate comparison, Literary and Linguistic Computing 21 (4) (2006) 463–476, Special Issue, J.Nerbonne & W.Kretzschmar (eds.), Progress in Dialectometry: Toward Explanation.

[11] J. Prokić, Identifying linguistic structure in a quantitative analysis of dialect pronunciation, in: Proc. of the ACL 2007 Student Research Workshop, Association for Computational Linguistics, Prague, 2007, pp. 61–66.

[12] I. Dhillon, Co-clustering documents and words using bipartite spectral graph partitioning, in: Proc. of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM New York, NY, USA, 2001, pp. 269–274.

[13] Y. Kluger, R. Basri, J. Chang, M. Gerstein, Spectral biclustering of microarray data: Coclustering genes and conditions, Genome Res 13 (4) (2003) 703–716.

[14] T. Goeman, J. Taeldeman, Fonologie en morfologie van de Nederlandse dialecten. Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten, Taal en Tongval 48 (1996) 38–59.

[15] B. van den Berg, Phonology & Morphology of Dutch & Frisian Dialects in 1.1 million transcriptions, Goeman-Taeldeman-Van Reenen project 1980-1995, Meertens Instituut Electronic Publications in Linguistics 3. Meertens Instituut (CD-ROM), Amsterdam, 2003.

[16] J. Taeldeman, G. Verleyen, De FAND: een kind van zijn tijd, Taal en Tongval 51 (1999) 217–240.

[17] G. De Schutter, B. van den Berg, T. Goeman, T. de Jong, Morfologische Atlas van de Nederlandse Dialecten (MAND) Deel 1, Amsterdam University Press, Meertens Instituut - KNAW, Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Amsterdam, 2005.

[18] M. Wieling, W. Heeringa, J. Nerbonne, An aggregate analysis of pronunciation in the Goeman-Taeldeman-Van Reenen-Project data, Taal en Tongval 59 (2007) 84–116.

[19] V. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, Doklady Akademii Nauk SSSR 163 (1965) 845–848.

[20] K. W. Church, P. Hanks, Word association norms, mutual information, and lexicography, Comput Linguist 16 (1) (1990) 22–29.

[21] M. Wieling, J. Prokić, J. Nerbonne, Evaluating the pairwise alignment of pronunciations, in: L. Borin, P. Lendvai (Eds.), Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education, 2009, pp. 26–34.

[22] F. Chung, Spectral graph theory, American Mathematical Society, 1997.

[23] J. Goossens, Inleiding tot de Nederlandse dialectologie, Wolters-Noordhoff, Groningen, 1977.

[24] W. Heeringa, Een andere indeling van de Limburgse dialecten, in: L. Heijenrath (Ed.), Veldeke Jaarboek 2007, Vereniging Veldeke Limburg, Roermond, 2008, pp. 94–104.

[25] J. Prokić, M. Wieling, J. Nerbonne, Multiple sequence alignments in linguistics, in: L. Borin, P. Lendvai (Eds.), Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education, 2009, pp. 18–25.

[26] W. Heeringa, B. Joseph, The relative divergence of dutch dialect pronunciations from their common source: An exploratory study, in: J. Nerbonne, T. M. Ellison, G. Kondrak (Eds.), Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, ACL, Stroudsburg, PA, 2007, pp. 31–39.

[27] Neuhochdeutsch-germanisches Wörterbuch, available at `http://www.koeblergerhard.de/germwbhinw.html`. Accessed: February 9, 2010.

[28] W. Heeringa, P. Kleiweg, C. Gooskens, J. Nerbonne, Evaluation of string distance algorithms for dialectology, in: J. Nerbonne, E. Hinrichs (Eds.), Linguistic Distances, ACL, Stroudsburg, PA, 2006, pp. 51–62.

[29] S. Madeira, A. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE/ACM Trans Comput Biol Bioinform 1 (1) (2004) 24–45.

[30] I. Dhillon, S. Mallela, D. Modha, Information-theoretic co-clustering, in: Proc. of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM New York, NY, USA, 2003, pp. 89–98.

[31] E. Schneider, Qualitative vs. quantitative methods of area delimitation in dialectology: A comparison based on lexical data from Georgia and Alabama, Journal of English Linguistics 21 (1988) 175–212.

Captions for figures

Figure 1: Distribution of GTRP localities including province names

Figure 2: Example of a bipartite graph of four varieties and three sound correspondences

Figure 3: Co-clustering varieties (y-axis) and segment substitutions (x-axis) in 2 (left), 3 (middle) and 4 (right) clusters

Figure 4: Geographic projection of varieties in 2 clusters (left), 3 clusters (middle) and 4 clusters (right)

Figure 2

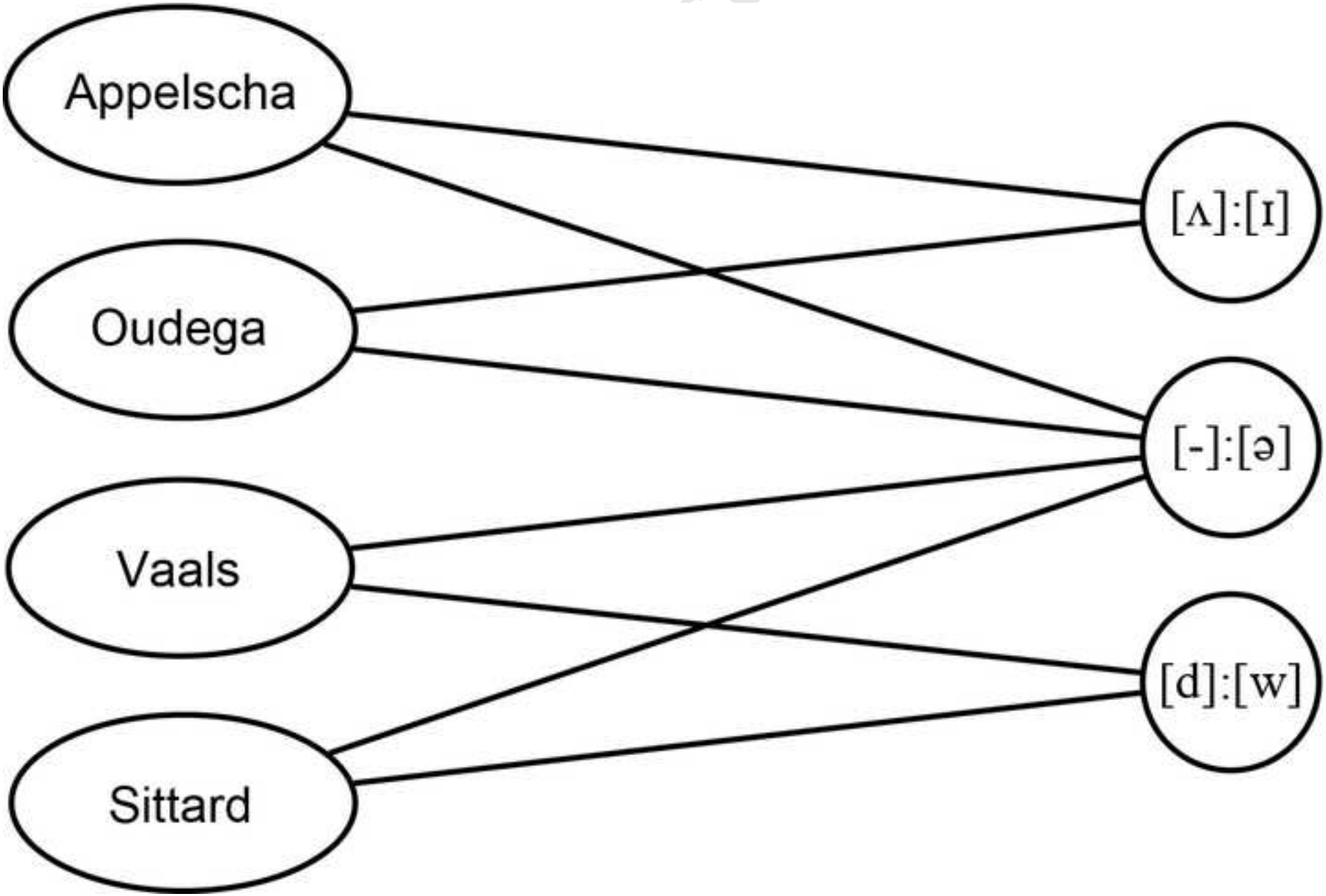


Figure 3

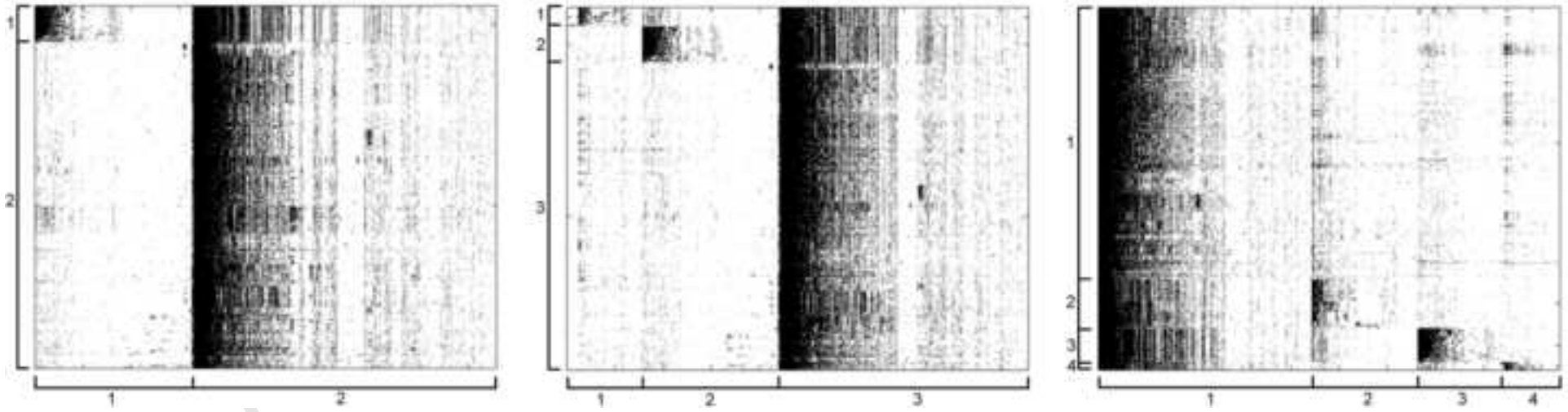


Figure 4