Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features

Martijn Wieling
University of Groningen
The Netherlands
m.b.wieling@rug.nl

John Nerbonne
University of Groningen
The Netherlands
j.nerbonne@rug.nl

Abstract

In this study we apply hierarchical spectral partitioning of bipartite graphs to a Dutch dialect dataset to cluster dialect varieties and determine the concomitant sound correspondences. An important advantage of this clustering method over other dialectometric methods is that the linguistic basis is *simultaneously* determined, bridging the gap between traditional and quantitative dialectology. Besides showing that the results of the hierarchical clustering improve over the flat spectral clustering method used in an earlier study (Wieling and Nerbonne, 2009), the values of the second singular vector used to generate the two-way clustering can be used to identify the most important sound correspondences for each cluster. This is an important advantage of the hierarchical method as it obviates the need for external methods to determine the most important sound correspondences for a geographical cluster.

1 Introduction

For almost forty years quantitative methods have been applied to the analysis of dialect variation (Séguy, 1973; Goebl, 1982; Nerbonne et al., 1999). Until recently, these methods focused mostly on identifying the most important dialectal groups using an aggregate analysis of the linguistic data.

One of these quantitative methods, clustering, has been applied frequently to dialect data, especially in an effort to compare computational analyses to traditional views on dialect areas (Davis and Houck, 1995; Clopper and Pisoni, 2004; Heeringa, 2004; Moisl and Jones, 2005; Mucha and Haimerl, 2005; Prokić and Nerbonne, 2009).

While viewing dialect differences at an aggregate level certainly gives a more comprehensive view than the analysis of a single subjectively selected feature, the aggregate approach has never fully convinced traditional linguists of its use as it fails to identify the linguistic distinctions among the identified groups. Recently, however, Wieling and Nerbonne (2009; 2010) answered this criticism by applying a promising graph-theoretic method, the spectral partitioning of bipartite graphs, to cluster varieties and simultaneously determine the linguistic basis of the clusters.

The spectral partitioning of bipartite graphs has been a popular method for the task of co-clustering since its introduction by Dhillon in 2001. Besides being used in the field of information retrieval for co-clustering words and documents (Dhillon, 2001), this method has also proven useful in the field of bioinformatics, successfully co-clustering genes and conditions (Kluger et al., 2003).

Wieling and Nerbonne (2009) used spectral partitioning of bipartite graphs to co-cluster dialect varieties and sound correspondences with respect to a set of reference pronunciations. They reported a fair geographical clustering of the varieties in addition to sensible sound correspondences. In a follow-up study, Wieling and Nerbonne (2010) developed an external method to rank the sound correspondences for each geographic cluster, which also conformed largely to the subjectively selected "interesting" sound correspondences in their earlier study (Wieling and Nerbonne, 2009).

In all the aforementioned studies, the spectral graph partitioning method was used to generate a flat clustering. However, Shi and Malik (2000) indicated that a hierarchical clustering obtained by repeatedly grouping in two clusters should be preferred over the flat clustering approach as approximation errors are reduced. More importantly, genealogical relationships between languages (or dialects) are generally expected to have a hierarchical structure due to the dynamics of language

Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pages 33–41,
Uppsala, Sweden, 16 July 2010.

Figure 1: Distribution of GTRP varieties including province names

change in which early changes result in separate varieties which then undergo subsequent changes independently (Jeffers and Lehiste, 1979).

In this study, we will apply the hierarchical spectral graph partitioning method to a Dutch dialect dataset. Besides comparing the results to the flat clustering obtained by Wieling and Nerbonne (2009), we will also show that identifying the most important sound correspondences is inherent to the method, alleviating the need for an external ranking method (e.g., see Wieling and Nerbonne, 2010).

While the current study applies the hierarchical clustering and (novel) ranking method to pronunciation data, we would also like to point out that these methods are not restricted to this type of data and can readily be applied to other domains such as information retrieval and bioinformatics where other spectral methods (e.g., principal component analysis) have already been applied successfully (e.g., see Furnas et al., 1988 and Jolicoeur and Mosimann, 1960).

2 Material

In this study, we use the same dataset as discussed by Wieling and Nerbonne (2009). In short, the Goeman-Taeldeman-Van Reenen-project data (GTRP; Goeman and Taeldeman, 1996; Van den Berg, 2003) is the most recent Dutch dialect dataset digitally available consisting of 1876 phonetically transcribed items for 613 dialect varieties in the Netherlands and Flanders. We focus on a subset of 562 words selected by Wieling et al. (2007) for all 424 Netherlandic varieties. We do not include the Belgian varieties, as the transcriptions did not use the same number of tokens as used for the Netherlandic transcriptions. The geographic distribution of the GTRP varieties including province names is shown in Figure 1.

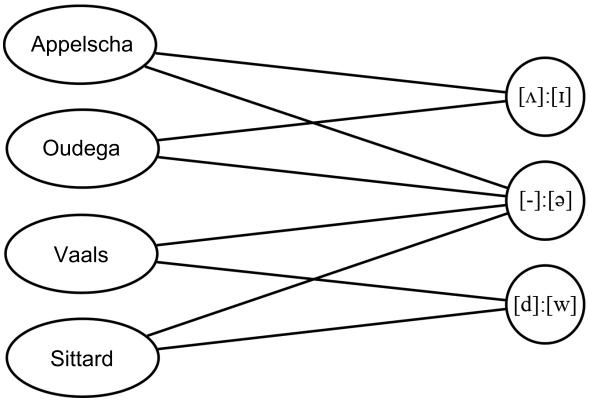

Figure 2: Example of a bipartite graph of varieties and sound correspondences

3 Methods

The spectral graph partitioning method we apply requires as input an undirected bipartite graph. A bipartite graph is a graph consisting of two sets of vertices where each edge connects a vertex from one set to a vertex in the other set. Vertices within a set are not connected. An example of a bipartite graph is shown in Figure 2. The vertices on the left side represent the varieties, while the vertices on the right side represent the sound correspondences (each individual sound is surrounded by a set of square brackets). An edge between a variety and a sound correspondence indicates that the sound correspondence occurs in that variety with respect to a specific reference variety.

As we are interested in clustering dialect varieties and detecting their underlying linguistic basis, our bipartite graph consists of dialect varieties and for each variety the presence of sound correspondences compared to a reference variety (indicated by an edge; see Figure 2). Because we do not have pronunciations of standard (or historical) Dutch, we use the pronunciations of the city Delft as our reference, since they are close to standard

Dutch (Wieling and Nerbonne, 2009) and allow a more straightforward interpretation of the sound correspondences than those of other varieties.

3.1 Obtaining sound correspondences

We obtain the sound correspondences by aligning the pronunciations of Delft against the pronunciations of all other dialect varieties using the Levenshtein algorithm (Levenshtein, 1965). The Levenshtein algorithm generates an alignment by minimizing the number of edit operations (insertions, deletions and substitutions) needed to transform one string into the other. For example, the Levenshtein distance between [bɪndən] and [bɛində], two Dutch dialect pronunciations 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, it is clear that the vowel [ɪ] in the alignment above can be aligned with either the vowel [ɛ] or the vowel [i]. To improve the initial alignments, we use an empirically derived segment distance table obtained by using the pointwise mutual information (PMI) procedure as introduced by Wieling et al. (2009).[1] They showed that applying the PMI procedure resulted in much better alignments than using several other alignment procedures.

The initial step of the PMI procedure consists of obtaining a starting set of alignments. In our case we obtain these 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. Subsequently, the substitution cost of every segment pair (a segment can also be a gap, representing an insertion or a deletion) can be 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.

In short, this procedure adapts the distance between two sound segments based on how likely it is that they are paired in the alignments. If two sounds are seen more (less) often together than we would expect based on their relative frequency in the dataset, their PMI score will be positive (negative). Higher scores indicate that segments tend to co-occur in correspondences more often, while lower scores indicate the opposite. New segment distances (i.e. segment substitution costs) are obtained by subtracting the PMI score from 0 and adding the maximum PMI score (to enforce that the minimum distance is 0). Based on the adapted segment distances we generate new alignments and we repeat this procedure until the alignments remain constant.

We extract the sound correspondences from the final alignments and represent the bipartite graph by a matrix $\boldsymbol{A}$ having 423 rows (all varieties, except Delft) and 957 columns (all occurring sound correspondences). We do not include frequency information in this matrix, but use binary values to indicate the presence (1) or absence (0) of a sound correspondence with respect to the reference pronunciation.[2] To reduce the effect of noise, we only

[1] The PMI procedure is implemented in the dialectometry package RUG/L04 which can be downloaded from http://www.let.rug.nl/ kleiweg/L04.

[2] We decided against using (the log) of the frequencies, as results showed that this approach gave too much weight to uninformative high-frequent substitutions of two identical sounds.

regard a sound correspondence as present in a variety when it occurs in at least three aligned pronunciations. Consequently, we reduce the number of sound correspondences (columns of $\boldsymbol{A}$) by more than half to 477.

3.2 Hierarchical spectral partitioning of bipartite graphs

Spectral graph theory is used to find the principal properties and structure of a graph from its graph spectrum (Chung, 1997). Wieling and Nerbonne (2009) used spectral partitioning of bipartite graphs as introduced by Dhillon (2001) to co-cluster varieties and sound correspondences, enabling them to obtain a geographical clustering with a simultaneously derived linguistic basis (i.e. the clustered sound correspondences). While Wieling and Nerbonne (2009) focused on the flat clustering approach, we will use the hierarchical approach by iteratively clustering in two groups. This approach is preferred by Shi and Malik (2000), because approximation errors are reduced compared to the flat clustering approach.

The hierarchical spectral partitioning algorithm, following Dhillon (2001), proceeds as follows:

1. Given the 423×477 variety-by-segment-correspondence matrix $\boldsymbol{A}$ as discussed previously, form

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

   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 $\boldsymbol{A_n}$

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

   and take the singular vectors $\boldsymbol{u_2}$ and $\boldsymbol{v_2}$

3. Compute $\boldsymbol{z_2} = \begin{bmatrix} \boldsymbol{D_1}^{-1/2} \boldsymbol{u_2} \\ \boldsymbol{D_2}^{-1/2} \boldsymbol{v_2} \end{bmatrix}$

4. Run the k-means algorithm on $\boldsymbol{z_2}$ to obtain the bipartitioning

5. Repeat steps 1 to 4 on both clusters separately to create the hierarchical clustering

The following example (taken from Wieling and Nerbonne, 2010) shows how we can co-cluster the graph of Figure 2 in two groups. The matrix representation of this graph is as follows:

	[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

The first step is to construct matrices $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$ which contain the total number of edges from every variety ($\boldsymbol{D_1}$) and every sound correspondence ($\boldsymbol{D_2}$) on the diagonal. Both matrices are shown below.

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

The normalized matrix $\boldsymbol{A_n}$ can be calculated using the formula displayed in step 1 of the hierarchical bipartitioning algorithm:

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

Applying the singular value decomposition to $\boldsymbol{A_n}$ yields:

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

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & .71 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

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

Finally, 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 $\boldsymbol{z_2}$:

$$\boldsymbol{z_2} = \begin{bmatrix} .35 & .35 & -.35 & -.35 & .5 & 0 & -.5 \end{bmatrix}^T$$

The first four values correspond with the places Appelscha, Oudega, Vaals and Sittard, while the

last three values correspond to the segment substitutions [ʌ]:[ɪ], [-]:[ə] and [d]:[w].

After running the k-means algorithm (with random initialization) on $\boldsymbol{z_2}$, the items are assigned to one of two clusters as follows:

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

This clustering shows that Appelscha and Oudega are grouped 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). Also, Vaals and Sittard are clustered together and linked to the clustered segment substitution [d]:[w] (cluster 2). The segment substitution [-]:[ə] (an insertion of [ə]) is actually not meaningful for the clustering of the varieties (as can be seen in $\boldsymbol{A}$), because the middle value of $\boldsymbol{V}^{\mathrm{T}}$ corresponding to this segment substitution equals 0. It could therefore just as likely be grouped cluster 2. Nevertheless, the k-means algorithm always assigns every item to one cluster.[3]

3.3 Determining the importance of sound correspondences

Wieling and Nerbonne (2010) introduced a *post hoc* method to rank each sound correspondence [a]:[b] based on the representativeness R in a cluster c_i (i.e. the proportion of varieties v in cluster c_i containing the sound correspondence):

$$R(\mathrm{a},\mathrm{b},c_i) = \frac{|v \text{ in } c_i \text{ containing [a]:[b]}|}{|v \text{ in } c_i|}$$

and the distinctiveness D (i.e. the number of varieties v within as opposed to outside cluster c_i containing the sound correspondence normalized by the relative size of the cluster):

$$D(\mathrm{a},\mathrm{b},c_i) = \frac{O(\mathrm{a},\mathrm{b},c_i) - S(c_i)}{1 - S(c_i)}$$

Where the relative occurrence O and the relative size S are given by:

$$O(\mathrm{a},\mathrm{b},c_i) = \frac{|v \text{ in } c_i \text{ containing [a]:[b]}|}{|v \text{ containing [a]:[b]}|}$$

$$S(c_i) = \frac{|v \text{ in } c_i|}{|\text{all } v\text{'s}|}$$

The importance I is then calculated by averaging the distinctiveness and representativeness:

$$I(\mathrm{a},\mathrm{b},c_i) = \frac{R(\mathrm{a},\mathrm{b},c_i) + D(\mathrm{a},\mathrm{b},c_i)}{2}$$

An extensive explanation of this method can be found in Wieling and Nerbonne (2010).

As we now only use a single singular vector to determine the partitioning (in contrast to the study of Wieling and Nerbonne, 2010 where they used multiple singular vectors to determine the flat clustering), we will investigate if the values of the singular vector $\boldsymbol{v_2}$ reveal information about the importance (as defined above) of the individual sound correspondences. We will evaluate these values by comparing them to the importance values on the basis of the representativeness and distinctiveness (Wieling and Nerbonne, 2010).

4 Results

In this section, we will report the results of applying the hierarchical spectral partitioning method to our Dutch dialect dataset. In addition, we will also compare the geographical clustering to the results obtained by Wieling and Nerbonne (2009).

We will only focus on the four main clusters each consisting of at least 10 varieties. While our method is able to detect smaller clusters in the data, we do not believe these to be stable. We confirmed this by applying three well-known distance-based clustering algorithms (i.e. UPGMA, WPGMA and Ward's Method; Prokić and Nerbonne, 2009) to our data which also only agreed on four main clusters. In addition, Wieling and Nerbonne (2009) reported reliable results on a maximum of 4 clusters.

4.1 Geographical results

Figure 3 shows a dendrogram visualizing the obtained hierarchy as well as a geographic visualization of the clustering. For comparison, Figure 4 shows the visualization of four clusters based on the flat clustering approach of Wieling and Nerbonne (2009).

It is clear that the geographical results of the hierarchical approach (Figure 3) are comparable to the results of the flat clustering approach (Figure 4) of Wieling and Nerbonne (2009).[4] How-

[3] Note that we could also have decided to drop this sound correspondence. However using our ranking approach (see Secion 3.3) already ensures that the uninformative sound correspondences are ranked very low.

[4] Note that the results of the flat clustering approach were based on all 957 sound correspondences. No noise-reducing frequency threshold was applied there, as this was reported to lead to poorer results (Wieling and Nerbonne, 2009).

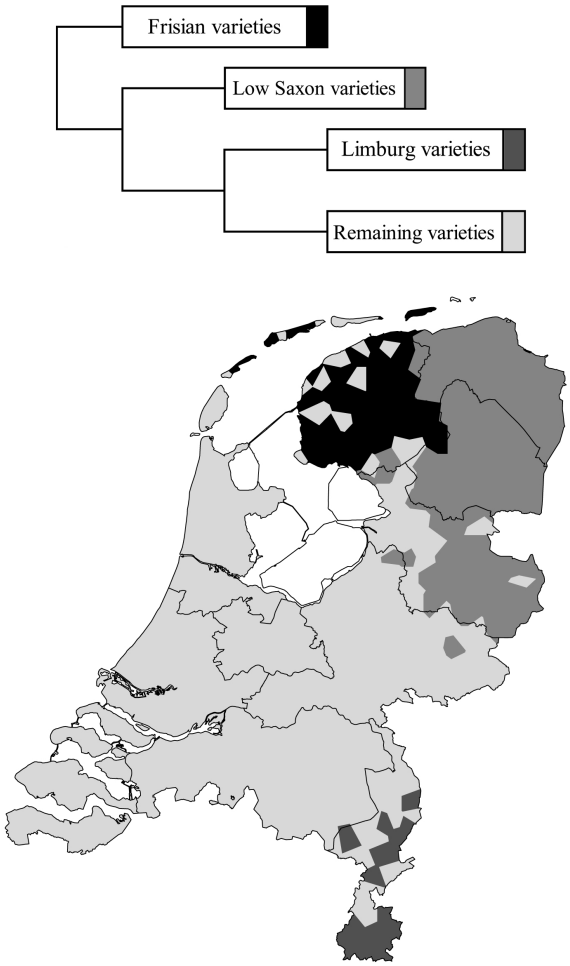

Figure 3: Geographic visualization of the clustering including dendrogram. The shades of grey in the dendrogram correspond with the map (e.g., the Limburg varieties can be found at the bottom-right).

ever, despite the Frisian area (top-left) being identical, we clearly observe that both the Low Saxon area (top-right) and the Limburg area (bottom-right) are larger in the map based on the hierarchical approach. As this better reflects the traditional Dutch dialect landscape (Heeringa, 2004), the hierarchical clustering method seems to be an improvement over the flat clustering method. Also the hierarchy corresponds largely with the one found by Heeringa (2004, Chapter 9), identifying Frisian, Limburg and Low Saxon as separate groups.

4.2 Most important sound correspondences

To see whether the values of the singular vector $\boldsymbol{v_2}$ can be used as a substitute for the external ranking method, we correlated the absolute values of the

Figure 4: Geographic visualization of the flat clustering reported in Wieling and Nerbonne (2009). The shades of grey are identical to the shades of grey in Figure 3.

singular vector with the importance values based on the distinctiveness and representativeness. For the sound correspondences of the Frisian area we obtained a high Spearman's rank correlation coefficient ρ of .92 ($p < .001$). For the Low Saxon area and the Limburg area we obtained similar values ($\rho = .87, p < .001$ and $\rho = .88, p < .001$, respectively). These results clearly show that the values of the second singular vector $\boldsymbol{v_2}$ can be used as a good substitute for the external ranking method.

Frisian area

The following table shows the five most important sound correspondences for the Frisian area.

Rank	1	2	3	4	5
Reference	-	[x]	[f]	[x]	[a]
Frisian	[ʃ]	[j]	-	[z]	[i]

While we have limited overlap (only [x]:[z]; occuring in e.g. *zeggen* 'say' Dutch [zɛxə], Frisian [sizə]) with the sound correspondences selected and discussed by Wieling and Nerbonne (2010) who used the flat clustering method without a frequency threshold (also causing some of the differences), we observe more overlap with the subjec-

tively selected sound correspondences in Wieling and Nerbonne (2009; [x]:[j] in e.g. *geld* 'money' Dutch [xɛlt], Frisian [jɪlt]; and [a]:[i] in e.g. *kaas* 'cheese' Dutch [kas], Frisian [tsis]). In addition, we detected two novel sound correspondences ([f]:[-] and [-]:[ʃ]).

We commonly find the correspondence [-]:[ʃ] in the infinitive form of verbs such as *wachten* 'wait' Dutch [waxtə], Frisian [waxtʃə]; *vechten* 'fight' Dutch [vɛxtə], Frisian [vɛxtʃə]; or *spuiten* 'spray' Dutch [spœʏtə], Frisian [spoʏtʃə], but it also appears e.g. in Dutch *tegen* 'against' [teixə], Frisian [tʃɪn]. The [f]:[-] correspondence is found in words like *sterven* 'die' standard Dutch [stɛrfə], Frisian [stɛrə].

Low Saxon area

The most important sound correspondences of the Low Saxon area are shown in the table below.

Rank	1	2	3	4	5
Reference	[k]	[v]	[ə]	[f]	[p]
Low Saxon	[ʔ]	[b]	[m]	[b]	[ʔ]

These sound correspondences overlap to a large extent with the most important sound correspondences identified and discussed by Wieling and Nerbonne (2010). The correspondence [k]:[ʔ] can be found in words like *planken* 'boards', Dutch [plɑŋkə], Low Saxon [plɑŋʔŋ̩], while the correspondence [v]:[b] is found in words like *bleven* 'remain' Dutch [blɛvən], Low Saxon [blibm̩]. The final overlapping correspondence [f]:[b] can be observed in words like *proeven* 'test' Dutch [prufə], Low Saxon [proybm̩].

The sound correspondence [ə]:[m] was discussed and selected by Wieling and Nerbonne (2009) as an interesting sound correspondence, occurring in words like *strepen* 'stripes' Dutch [strepə], Low Saxon [strepm̩].

The new correspondence [p]:[ʔ] occurs in words such as *lampen* 'lamps' standard Dutch [lampə], Aduard (Low Saxon) [lamʔm̩], but also postvocalically, as in *gapen* 'yawn', standard Dutch [xapə], Aduard (Low Saxon) [xoʔm̩]. It is obviously related to the [k]:[ʔ] correspondence discussed above.

Limburg area

The most important sound correspondences for the Limburg area are displayed in the table below.

Rank	1	2	3	4	5
Reference	[r]	[s]	[o]	[n]	[r]
Limburg	[x]	[ʒ]	-	[x]	[ʀ]

For this area, we observe limited overlap with the most important sound correspondences based on distinctiveness and representativeness (Wieling and Nerbonne, 2010; only [n]:[x] overlaps, occurring in words like *kleden* 'cloths' Dutch [kledən], Limburg [klɛɪdəx]), as well as with the subjectively selected interesting sound correspondences (Wieling and Nerbonne, 2009; only [r]:[ʀ] overlaps, which occurs in words like *breken* 'to break' Dutch [brekə], Limburg [bʀɛkə]).

The sound correspondence [o]:[-] can be found in *wonen* 'living', pronounced [wounə] in our reference variety Delft and [wunə] in Limburg. As the standard Dutch pronunciation is actually [wonə], this correspondence is caused by the choice of our reference variety, which is unfortunately not identical to standard Dutch.

The other two sound correspondences are more informative. The sound correspondence [r]:[x] can be found in words like *vuur* 'fire' Dutch [fyr], Limburg [vyəx] and is similar to the sound correspondence [r]:[ʀ] discussed above. The other correspondence [s]:[ʒ] occurs when comparing the standard-like Delft variety to Limburg varietes in words such as *zwijgen* 'to be silent' [sweixə], Limburg [ʒwiɣə]; or *zwemmen* 'swim' [swɛmə], Limburg [ʒwɵmə].

Hierarchical versus flat clustering

In general, then, the sound correspondences uncovered by the hierarchical version of the spectral clustering technique turn out to be at least as interesting as those uncovered by the flat clustering, which leads us to regard the hierarchical clustering technique as defensible in this respect. Since dialectologists are convinced that dialect areas are organized hierarchically, we are naturally inclined toward hierarchical clustering techniques as well. We note additionally that the using the values of the second singular vector is an adequate substitution of the external ranking method based on distinctiveness and representativeness, which means that the present paper also marks a step forward in simplifying the methodology.

5 Discussion

In this study we showed that using hierarchical spectral partitioning of bipartite graphs results

in an improved geographical clustering over the flat partitioning method and also results in sensible concomitant sound correspondences. One of the reasons for the improvement of the geographical clustering could be the approximation errors which arise when going from the real valued solution to the discrete valued solution, and which increase with every additional singular vector used (Shi and Malik, 2000).

In addition, we showed that using the values of the second singular vector obviates the need for an external ranking method (e.g., see Wieling and Nerbonne, 2010) to identify the most important sound correspondences.

Since the spectral partitioning of bipartite graphs appears to be identifying significant (representative and distinctive) correspondences well – both in the flat clustering design and in the (present) hierarchical scheme, several further opportunities become worthy of exploration. First, we might ask if we can automatically identify a threshold of significance for such correspondences, as to-date we have only sought to verify significance, not to exclude marginally significant elements. Second, while we have applied the technique exclusively to data for which the correspondence consists of a comparison of dialect data to (near) standard data, the analysis of historical data, in which varieties are compared to an earlier form, is within reach. As the first step, we should wish to compare data to a well-established historical predecessor as further steps might require genuine reconstruction, still beyond anyone's reach (as far as we know). Third, the technique would be more generally applicable if it did not require agreeing on a standard, or pole of comparison. This sounds difficult, but multi-alignment techniques might bring it within reach (Prokić et al., 2009).

It is intriguing to note that Nerbonne (in press) found only sporadic correspondences using factor analysis on data which incorporated frequency of correspondence, and we have likewise found frequency-weighted data less successful as a basis for spectral bipartite clustering. Shackleton (2007), Wieling and Nerbonne (2010) and the current paper are more successful using data which lacks information about the frequency of occurrence of sounds and/or sound correspondences. The question arises as to whether this is general and why this is so. Is it due to the skewness of frequency distributions, in which a suitable normalization might be attempted? Or is it simply more straightforward to focus on the absolute presence or absence of a sound or sound correspondence?

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

Finally, while we only tested this method on a single dataset, it would be interesting to see if our results and conclusions also hold when applied to more and different datasets. We also realize that the evaluation in this study is rather qualitative, but we intend to develop more quantitative evaluation methods for future studies.

Acknowledgements

We thank Gertjan van Noord and Tim Van de Cruys for their comments during a presentation about the flat spectral graph partitioning method, which instigated the search for an inherent ranking method.

References

Fan Chung. 1997. *Spectral graph theory*. American Mathematical Society.

Cynthia G. Clopper and David B. Pisoni. 2004. Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, 32(1):111–140.

L.M. Davis and C.L. Houck. 1995. What Determines a Dialect Area? Evidence from the Linguistic Atlas of the Upper Midwest. *American Speech*, 70(4):371–386.

Inderjit Dhillon. 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*, pages 269–274. ACM New York, NY, USA.

George Furnas, Scott Deerwester, Susan Dumais, Thomas Landauer, Richard Harshman, Lynn Streeter, and Karen Lochbaum. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–480. ACM.

Hans Goebl. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*.

Österreichische Akademie der Wissenschaften, Wien.

Ton Goeman and Johan Taeldeman. 1996. Fonologie en morfologie van de Nederlandse dialecten. Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval*, 48:38–59.

Wilbert Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph.D. thesis, Rijksuniversiteit Groningen.

Robert Jeffers and Ilse Lehiste. 1979. *Principles and methods for historical linguistics*. MIT Press, Cambridge.

Pierre Jolicoeur and James E. Mosimann. 1960. Size and shape variation in the painted turtle. A principal component analysis. *Growth*, 24:339–354.

Yuval Kluger, Ronen Basri, Joseph Chang, and Mark Gerstein. 2003. Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research*, 13(4):703–716.

Vladimir Levenshtein. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163:845–848.

Hermann Moisl and Val Jones. 2005. Cluster analysis of the newcastle electronic corpus of tyneside english: A comparison of methods. *Literary and Linguistic Computing*, 20(supp.):125–146.

Hans-Joachim Mucha and Edgard Haimerl. 2005. Automatic validation of hierarchical cluster analysis with application in dialectometry. In Claus Weihs and Wolgang Gaul, editors, *Classification—the Ubiquitous Challenge. Proc. of the 28th Meeting of the Gesellschaft für Klassifikation, Dortmund, March 9–11, 2004*, pages 513–520, Berlin. Springer.

John Nerbonne, Wilbert Heeringa, and Peter Kleiweg. 1999. Edit distance and dialect proximity. In David Sankoff and Joseph Kruskal, editors, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.*, pages v–xv. CSLI, Stanford, CA.

John Nerbonne. in press. Various Variation Aggregates in the LAMSAS South. In C. Davis and M. Picone, editors, *Language Variety in the South III*. University of Alabama Press, Tuscaloosa.

Jelena Prokić and John Nerbonne. 2009. Recognizing groups among dialects. In John Nerbonne, Charlotte Gooskens, Sebastian Kurschner, and Rene van Bezooijen, editors, *International Journal of Humanities and Arts Computing, special issue on Language Variation*.

Jelena Prokić, Martijn Wieling, and John Nerbonne. 2009. Multiple sequence alignments in linguistics. In Lars Borin and Piroska Lendvai, editors, *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 18–25.

Jean Séguy. 1973. La dialectométrie dans l'atlas linguistique de gascogne. *Revue de Linguistique Romane*, 37(145):1–24.

Robert G. Shackleton, Jr. 2007. Phonetic variation in the traditional english dialects. *Journal of English Linguistics*, 35(1):30–102.

Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905.

Boudewijn van den Berg. 2003. *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.

Martijn Wieling and John Nerbonne. 2009. Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology. In Monojit Choudhury, Samer Hassan, Animesh Mukherjee, and Smaranda Muresan, editors, *Proc. of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 26–34.

Martijn Wieling and John Nerbonne. 2010. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language*. Accepted to appear in a special issue on network models of social and cognitive dynamics of language.

Martijn Wieling, Wilbert Heeringa, and John Nerbonne. 2007. An aggregate analysis of pronunciation in the Goeman-Taeldeman-Van Reenen-Project data. *Taal en Tongval*, 59:84–116.

Martijn Wieling, Jelena Prokić, and John Nerbonne. 2009. Evaluating the pairwise alignment of pronunciations. In Lars Borin and Piroska Lendvai, editors, *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 26–34.