Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology

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 used bipartite spectral graph partitioning to simultaneously cluster varieties and sound correspondences in Dutch dialect data. While clustering geographical varieties with respect to their pronunciation is not new, the simultaneous identification of the sound correspondences giving rise to the geographical clustering presents a novel opportunity in dialectometry. Earlier methods aggregated sound differences and clustered on the basis of aggregate differences. The determination of the significant sound correspondences which co-varied with cluster membership was carried out on a *post hoc* basis. Bipartite spectral graph clustering simultaneously seeks groups of individual sound correspondences which are associated, even while seeking groups of sites which share sound correspondences. We show that the application of this method results in clear and sensible geographical groupings and discuss the concomitant sound correspondences.

1 Introduction

Exact methods have been applied successfully to the analysis of dialect variation for over three decades (Séguy, 1973; Goebl, 1982; Nerbonne et al., 1999), 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 material (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) (Nerbonne, 2009).

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 (2002) 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 (2004: 268–270) calculated which words correlated best with the first, second and third dimensions of an MDS analysis of aggregate pronunciation differences. Shackleton (2004) 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 (2006) 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 (2007) 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 (2001) 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

Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, ACL-IJCNLP 2009, pages 14–22,
Suntec, Singapore, 7 August 2009. ©2009 ACL and AFNLP

his study, Dhillon (2001) also demonstrated that his algorithm worked well on real world examples.

The usefulness of this approach is not only limited to clustering documents and words simultaneously. For example, Kluger et al. (2003) used a somewhat adapted bipartite spectral graph partitioning approach to successfully cluster microarray data simultaneously in clusters of genes and conditions.

In summary, the contribution of this paper 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. It is worth noting that, in isolating the linguistic basis of varietal affinities, we thereby hope to contribute technically to the study of how cognitive and social dynamics interact in language variation. Although we shall not pursue this explicitly in the present paper, our idea is very simple. The geographic signal in the data is a reflection of the social dynamics, where geographic distance is the rough operationalization of social contact. In fact, dialectometry is already successful in studying this. We apply techniques to extract (social) associations among varieties and (linguistic) associations among the speech habits which the similar varieties share. The latter, linguistic associations are candidates for cognitive explanation. Although this paper cannot pursue the cognitive explanation, it will provide the material which a cognitive account might seek to explain.

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, both the alignment technique used to obtain sound correspondences, as well as the bipartite spectral graph partitioning we used to simultaneously seek affinities in varieties as well as affinities in sound correspondences. 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; Goeman and Taeldeman, 1996; Van den Berg, 2003). 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 in Taeldeman and Verleyen (1999).

Because the GTRP was compiled with a view to documenting both phonological and morphological variation (De Schutter et al., 2005) and our purpose here is the analysis of sound correspondences, we ignore many items of the GTRP. We use the same 562 item subset as introduced and discussed in depth in Wieling et al. (2007). In short, the 1876 item word list was filtered by selecting only single word items, plural nouns (the singular form was 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 sound correspondences. In all varieties the same lexeme was used for a single item.

Because the GTRP transcriptions of Belgian varieties are fundamentally different from transcriptions of Netherlandic varieties (Wieling et al., 2007), 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.

3 Methods

To obtain the clearest signal of varietal differences in sound correspondences, 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. There are also no pronunciations in standard Dutch in the GTRP and transcribing the standard Dutch pronunciations ourselves would likely have introduced between-transcriber inconsistencies. Heeringa (2004: 274–276) identified pronunciations in the variety of Haarlem as being the closest to standard Dutch.

Figure 1: Distribution of GTRP localities including province names

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 (Levenshtein, 1965).

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 [lɛɪkən] and [likhən], two Dutch variants of the word 'seem', is 4:

lɛɪkən	delete ɛ	1
lɪkən	subst. i/ɪ	1
likən	insert h	1
likhən	subst. ɵ/ə	1
likhən		
		4

The corresponding alignment is:

l	ɛ	ɪ	k		ə	n
l		i	k	h	ɵ	n
	1	1		1	1	

When all edit operations have the same cost, multiple alignments yield a Levenshtein distance of 4 (i.e. by aligning the [i] with the [ɛ] and/or by aligning the [ə] with the [h]). To obtain only the best alignments we used an adaptation of the Levenshtein algorithm which uses automatically generated segment substitution costs. This procedure was proposed and described in detail by Wieling et al. (2009) and resulted in significantly better individual alignments than using the regular Levenshtein algorithm.

In brief, 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. I.e. if two segments are aligned more often than would be expected on the basis of their frequency in the dataset, the cost of substituting the two symbols is set relatively low; otherwise it is set relatively high. After the new segment substitution costs have been calculated, the strings are aligned again based on the new segment substitution costs. The previous two steps are then iterated 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 therefore obtain an $m \times n$ 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, 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. 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.

Spectral graph theory is used to find the principal properties and structure of a graph from its graph spectrum (Chung, 1997). Dhillon (2001) 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. In similar fashion, we would like to obtain a clustering of varieties and corresponding segment substitutions. We therefore apply the multipartitioning algorithm introduced by Dhillon (2001) to find k clusters:

1. Given the $m \times n$ variety-by-segment-substitution 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 $\boldsymbol{A_n}$

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

   and take the $l = \lceil \log_2 k \rceil$ singular vectors, $\boldsymbol{u_2}, \ldots, \boldsymbol{u_{l+1}}$ and $\boldsymbol{v_2}, \ldots, \boldsymbol{v_{l+1}}$

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

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-substitution matrix $\boldsymbol{A}$ in $k = 2$ groups.

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

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

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

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

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

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

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

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

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

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

After running the k-means algorithm on $\boldsymbol{Z}$, the items are assigned to one of two clusters as follows:

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

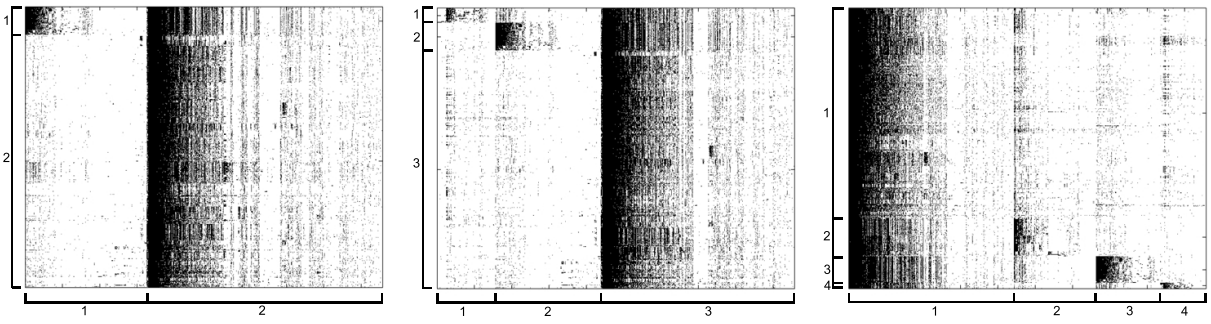

Figure 2: Visualizations of co-clustering varieties (y-axis) and segments substitutions (x-axis) in 2 (left), 3 (middle) and 4 (right) clusters

The clustering shows that Appelscha and Oudega are clustered together and linked to the clustered segment substitution of [ʌ]:[ɪ] (cluster 2). Similarly, Vaals and Sittard are clustered together and linked to the clustered segment substitutions [d]:[w] and [-]:[ə] (cluster 1). 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 bottom value of the second column of $\boldsymbol{V}$ 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.

In the following section we will report the results on clustering in two, three and four groups.[1]

4 Results

After running the multipartitioning algorithm[2] we obtained a two-way clustering in k clusters of varieties and segment substitutions. Figure 2 tries to visualize 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 3.

In the following subsections we will discuss the most important geographical clusters together with their simultaneously derived sound correspondences. For brevity, we will only focus on explaining a few derived sound correspondences for the most important geographical groups. 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 only occur in that particular cluster of varieties. This can also be observed in Figure 2, where sound correspondences in a particular cluster of varieties also appear in other clusters (but less dense).[3]

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 (2004: 227–229) 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 3) correspond to the Frisian cities which are generally found to deviate from the rest of the Frisian language area (Heeringa, 2004: 235–241).

[1] 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 the majority of information contained in the alignments it is more likely that this causes a decrease in the number of clusters we can reliably detect.

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

[3] In this study, we did not focus on identifying the most important sound correspondences in each cluster. See the Discussion section for a possible approach to rank the sound correspondences.

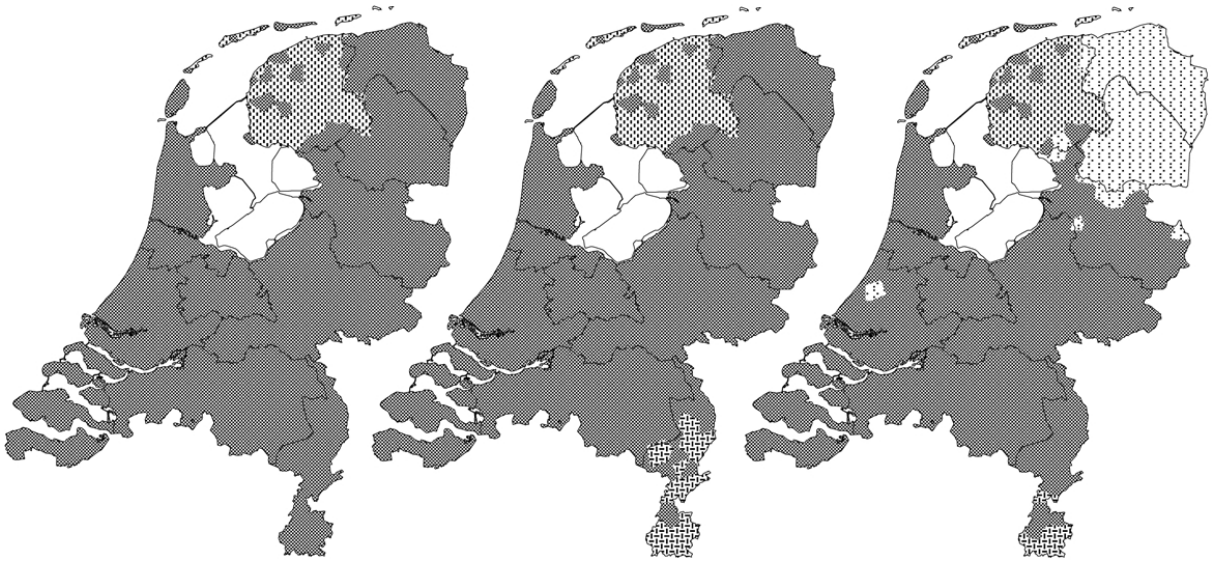

Figure 3: Clustering of varieties in 2 clusters (left), 3 clusters (middle) and 4 clusters (right)

A few interesting sound correspondences between the reference variety (Delft) and the Frisian area are displayed in the following table and discussed below.

Reference	[ʌ]	[ʌ]	[a]	[o]	[u]	[x]	[x]	[r]
Frisian	[ɪ]	[i]	[i]	[ɛ]	[ɛ]	[j]	[z]	[x]

In the table 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] (pronunciation 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] (Franeker) 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'. Frisian has lost the final [x], and Dutch has either lost a final [r] or experienced metathesis. 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.

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 (2004: 227–229) 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 important segment correspondences for Limburg are displayed in the following table and discussed below.

Reference	[r]	[r]	[k]	[n]	[n]	[w]
Limburg	[ʀ]	[ʁ]	[x]	[ʀ]	[ʁ]	[f]

Southern 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 clus-

tering 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 reflected 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 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.

The Low Saxon area

Finally, the division in 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 (see Heeringa, 2004), 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. It is nonetheless the case that Groningen and Drenthe normally are seen to form a separate northern subgroup within Low Saxon (Heeringa, 2004: 227–229).

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

Reference	[ə]	[ə]	[ə]	[-]	[a]
Low Saxon	[m]	[ŋ]	[ɴ]	[ʔ]	[e]

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].

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 [e] as in *kaas* 'cheese', [kes] in Gasselte, Hooghalen and Norg.

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 (2001) gave sensible clustering results in the geographical domain as well as for the concomitant linguistic basis.

As mentioned above, we did not have transcriptions of standard Dutch, but instead we used transcriptions of a variety (Delft) close to the standard langueage. 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 (see Prokić, 2009) all pronunciations of a single word and use the most frequent phonetic symbol at each position in the reconstructed word. 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].

In this study we did not investigate methods to identify the most important sound correspondences. A possible option would be to create a ranking procedure based on the uniqueness of the sound correspondences in a cluster. I.e. the sound

correspondence is given a high importance when it only occurs in the designated cluster, while the importance goes down when it also occurs in other clusters).

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.

As this study was the first to investigate the effectiveness of a co-clustering approach in dialectometry, we focused on the original bipartite spectral graph partitioning algorithm (Dhillon, 2001). Investigating other approaches such as biclustering algorithms for biology (Madeira and Oliveira, 2004) or an information-theoretic co-clustering approach (Dhillon et al., 2003) 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 (Goebl, 1982), 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 not addressed the interaction between cognitive and social dynamics directly, but we feel it has improved our vantage point for understanding this interaction. 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. Other dialectometric techniques have done this as well.

Linguists have rightly complained, however, that the linguistic factors have been neglected in dialectometry (Schneider, 1988:176). The current approach does not offer a theoretical framework to explain cognitive effects such as phonemes corresponding across many words, but does enumerate them clearly. This paper has shown that bipartite graph clustering can detect the linguistic basis of dialectal affinity. If deeper cognitive constraints are reflected in that basis, then we are now in an improved position to detect them.

Acknowledgments

We would like to 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 Jelena Prokić and the anonymous reviewers for their helpful comments on an earlier version of this paper.

References

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

Georges De Schutter, Boudewijn van den Berg, Ton Goeman, and Thera de Jong. 2005. *Morfologische Atlas van de Nederlandse Dialecten (MAND) Deel 1*. Amsterdam University Press, Meertens Instituut - KNAW, Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Amsterdam.

Inderjit Dhillon, Subramanyam Mallela, and Dharmendra Modha. 2003. Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98. ACM New York, NY, USA.

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.

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.

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.

Grzegorz Kondrak. 2002. Determining recurrent sound correspondences by inducing translation

models. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002)*, pages 488–494, Taipei. COLING.

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

Sara Madeira and Arlindo Oliveira. 2004. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45.

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

John Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

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.

Jelena Prokić. 2007. Identifying linguistic structure in a quantitative analysis of dialect pronunciation. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 61–66, Prague, June. Association for Computational Linguistics.

Edgar Schneider. 1988. Qualitative vs. quantitative methods of area delimitation in dialectology: A comparison based on lexical data from georgia and alabama. *Journal of English Linguistics*, 21:175–212.

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

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

Johan Taeldeman and Geert Verleyen. 1999. De FAND: een kind van zijn tijd. *Taal en Tongval*, 51:217–240.

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, 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.