# REASSESSING ENGLISH DIALECT VARIATION: A BIPARTITE SPECTRAL GRAPH PARTITIONING APPROACH

**Norova Mavluda Fayzulloyevna**
***associate professor of the department of Uzbek Language and Literature, Russian and English Languages at Bukhara State Medical Institute named after Abu Ali ibn Sino***

**Abstract.** This study evaluates hierarchical bipartite spectral graph partitioning (BiSGP) as a method for analyzing regional linguistic variation in England. Building on longstanding dialectological research and more recent advances in dialectometry, we apply BiSGP to the phonological dataset compiled by Shackleton (2007) from the *Survey of English Dialects* and the *Structural Atlas of the English Dialects*. Comparison with Shackleton's cluster analysis and varimax-rotated PCA demonstrates broad agreement on major dialect zones but also highlights differences: BiSGP occasionally merges or subdivides regions differently and uniquely identifies region-specific variant combinations, most notably in the "Potteries" area. The findings show that BiSGP provides a complementary perspective on dialect structure by emphasizing co-occurrence patterns across localities and variants, offering an empirically grounded alternative to traditional clustering and dimensionality-reduction techniques.

**Keywords:** dialectology; dialectometry; bipartite spectral graph partitioning; English dialects; linguistic variation; cluster analysis; principal component analysis; Survey of English Dialects.

# ПЕРЕОЦЕНКА ВАРИАТИВНОСТИ АНГЛИЙСКИХ ДИАЛЕКТОВ: МЕТОД ДВУДОЛЬНОГО СПЕКТРАЛЬНОГО РАЗБИЕНИЯ ГРАФА

**Аннотация.** В данном исследовании рассматривается и оценивается иерархический метод двудольного спектрального разбиения графа (BiSGP) как инструмент анализа региональной языковой вариативности в Англии. Опираясь на долгую традицию диалектологических исследований и современные достижения диалектометрии, мы применяем BiSGP к фонологическому датасету, составленному Шаклтоном (2007) на основе *Survey of English Dialects* и *Structural Atlas of the English Dialects*. Сравнение с кластерным анализом и варимакс-вращением главных компонент, использованными Шаклтоном, показывает значительное совпадение в выделении основных диалектных зон, но также выявляет и расхождения: BiSGP иногда по-другому объединяет или разделяет регионы и обнаруживает уникальные сочетания вариантов, характерные для отдельных локальных зон, особенно заметные в области «Potteries». Полученные результаты демонстрируют, что BiSGP предлагает дополнительный взгляд на диалектную структуру, подчёркивая закономерности совместного появления вариантов в разных локальностях и обеспечивая эмпирически обоснованную альтернативу традиционным методам кластеризации и методам снижения размерности.

**Ключевые слова:** диалектология; диалектометрия; двудольное спектральное разбиение графа; английские диалекты; языковая вариативность; кластерный анализ; метод главных компонент; Survey of English Dialects.

**Introduction.** A substantial amount of linguistic variation is shaped by geography, resulting in regional dialects that linguists have examined for more than a hundred years. By the 1970s, concerns about dialectology's strong focus on fine-grained detail led to the development of dialectometry, a field that systematizes analytical procedures and reduces the need for manually selecting linguistic features.

According to Nerbonne (2009), dialectometry's success stems from its emphasis on measuring *overall* similarity or difference, thereby enhancing the geographic patterns present in complex and sometimes contradictory linguistic data.

Although dialectometry has been politely received, some researchers worry that its focus on aggregated measures downplays the detailed linguistic information that could clarify underlying structure in variation. This has motivated several recent efforts to complement aggregate dialectometric approaches with methods for identifying linguistic variables that cluster geographically and methods that extract especially influential linguistic features tied to broad differentiation patterns.

Ruette and Speelman (submitted) contributed to this work by applying a form of three-way multidimensional scaling — individual differences scaling — to variationist research. Much like standard two-way multidimensional scaling (Nerbonne, 2010), it groups similar varieties using a distance matrix, while also exposing structure among the variables themselves.

Grieve et al. (2011) examined lexical variation in a large written English dataset. They applied spatial autocorrelation to identify significant geographic patterns across 40 lexical alternation variables, then used factor analysis to determine the contribution of each variable to the resulting factors, which broadly represent geographic areas. Cluster analysis was then applied to the factor scores to produce regional groupings.

Shackleton (2007) used cluster analysis and principal component analysis (PCA) to identify linguistic variables that commonly co-occur across multiple locations. For instance, if localities raising /æ/ to [ɛ] also tend to raise /e/ to [eɪ], cluster analysis should group these areas together, and PCA should identify a component capturing this shared pattern. Shackleton identified several such clusters and components, which we later revisit.

From a dialectometric standpoint, the BiSGP method is appealing because it simultaneously highlights features and localities. However, like all data-driven grouping methods, its usefulness must be demonstrated empirically. Here, we test its broader applicability by applying it to Shackleton's (2007) dataset and comparing the results with those obtained through cluster analysis and PCA.

**Material.** Our study uses the dataset presented in Shackleton (2007), which is based primarily on Anderson's (1987) *A Structural Atlas of the English Dialects* (SAED). The SAED comprises over 100 maps that chart the geographic distribution and frequency of phonetic variants appearing in word groups from the *Survey of English Dialects* (SED; Orton et al., 1962–1971). The SED provides the most comprehensive record of traditional dialect forms used in 313 rural English localities in the mid-20th century.

From the SAED maps, more than 400 responses were categorized into 39 groups. Each group of words contains one or more segments believed to have had a uniform pronunciation in the "standard" Middle English dialect of southeastern England's Home Counties. These segments include the full inventory of Middle English short and long vowels, diphthongs, and the small set of consonants that vary across English dialects.

For each Middle English target pronunciation, speakers in the SED may use several modern forms, and different words in the same group may be pronounced differently within a single locality. As a result, the dataset records the usage frequencies of 199 distinct variant realizations of the 39 underlying phonemes.

For example, one group includes words such as *root* and *tooth*, which shared the Middle English vowel /o:/. Multiple maps correspond to that group, each representing the distribution of a particular modern reflex. One shows how often /o:/ has shifted to [u:], another how often it has become [ʏ:], and so on. Throughout this

chapter, we use /x/ to represent the Middle English form and [x] to represent the SED variants. A full list of variants can be found in Shackleton (2010, pp. 180–186).

**Methods.** We represent the bipartite graph as a locality × variant matrix, where each cell contains the relative frequency (between 0 and 1) of a variant in a locality, following Shackleton (2007). To give all variants comparable influence, each column was rescaled so that values ranged from 0 to 1 by dividing every frequency by the highest frequency observed for that variant. This procedure may give proportionally greater weight to variants that are regionally distinctive but relatively rare.

***Distinctiveness* and *representativeness*.** Although these are typically averaged, they can also be weighted differently. When a dataset contains many variants that appear nearly everywhere, representativeness becomes uninformatively high; in such cases, distinctiveness should receive greater weight. Conversely, when many variants occur in only a few locations, representativeness should be weighted more strongly. Because our dataset contains many frequent variants, distinctiveness was weighted twice as heavily as representativeness.

**Results.** When applied to Shackleton's (2007) dataset, the BiSGP method first separates English dialects into northern and southern regions along a boundary that closely parallels Shackleton's division between northern, Midlands, and southern dialect areas. Among the approximately 70 high-scoring southern variants, a few are broadly distributed — often reflecting shifts or lengthening of Middle English short vowels. Most, however, are associated with subregional patterns, such as:

- up-gliding diphthongization of Middle English long vowels (e.g., [lɛɪn]/[læɪn] for *lane*) in the southeast,
- fricative voicing and rhoticity retention (e.g., [vɑːrm] for *farm*) in the southwest, and
- front vowel advancement (e.g., [nʉːn] for *noon*) in Devon.

The roughly 50 most characteristic northern variants show similar mixed behavior: some reflect conservative continuations of Middle English short vowels (e.g., [man] for *man*), while others include more localized in-gliding diphthongs (e.g., [liən] for *lane*), limited rhoticity, and fronting patterns such as [bøːn] for *bone* in the far north.

**Comparison to traditional cluster analysis.** BiSGP results can be directly compared with those from Shackleton's (2007) cluster analysis, highlighting the different strengths of the two methods. Unlike BiSGP, cluster analysis uses measures of overall similarity in variant usage among localities, without explicitly balancing representativeness and distinctiveness.

Shackleton applied several clustering procedures to the dataset, then merged their results into a site × site matrix of mean cophenetic distances (i.e., dendrogram distances). He performed multidimensional scaling on this matrix to reduce the data to a small set of dimensions that capture its major structural patterns. For visualization, three dimensions were mapped onto RGB color values.

As mentioned earlier, the division between northern and southern dialects resembles Shackleton's results, though BiSGP assigns a few Central Midlands localities to the northern group. Both methods identify nearly identical southeastern and southwestern regions and agree closely on the Northumberland region, aside from BiSGP isolating a single Cumberland locality due to its distinctive rhotic trill [r]. These peripheral regions stand out because they are characterized by coherent geographic distributions of particular variants—such as rhoticity and aspirates in the far north, fricative voicing in the southwest, vowel fronting in Devon, and strong up-gliding diphthongs in the southeast—making them relatively easy to detect across methods.

**Comparison to principal component analysis.** The BiSGP findings can also be compared with Shackleton's varimax-rotated PCA, which highlights

complementary strengths. While BiSGP emphasizes the representativeness and distinctiveness of variant frequencies within regions, PCA identifies clusters of variants that strongly co-occur or mutually exclude one another, grouping these into principal components.

Each component typically forms two poles: one with high positive loadings for one set of co-occurring variants, and another with high negative loadings for a mutually exclusive set. Varimax rotation sharpens these groupings by increasing loadings on the most influential variants, often making the linguistic interpretation clearer. Localities receive scores indicating how strongly each component is represented in their speech. When groups of localities exhibit similar scores with clear geographic coherence, they may correspond to identifiable dialect regions. PCA identifies variant groupings for around twelve regions of England, capturing about half of the total variation in the dataset.

Although PCA can provide a relatively objective description of some traditional dialect regions, it does not fully divide England into a complete set of regions. It also tends to identify variants that are unique to small areas or variants that are not exclusive to the hypothesized region, and few localities exhibit most of the variants associated with any one component.

**Discussion.** Hierarchical bipartite spectral graph partitioning offers a valuable complement to other dialectological methods because it identifies groups of localities that are linguistically similar at the same time as it identifies sets of variants that frequently co-occur. This creates an inductive framework in which linguistic and geographic patterns reinforce one another. After clustering, we identified the variants most strongly associated with each region by combining measures of distinctiveness (their frequency inside vs. outside the region) and representativeness (their frequency within the region).

This dual-focus approach contrasts with one-dimensional clustering, which groups localities solely on the basis of similar overall usage patterns, and with PCA, which reveals correlated sets of variants. Using the English dialect dataset examined in this chapter, BiSGP isolates dialect regions that broadly align with those identified through cluster analysis and PCA and highlights sets of distinctive variants that also largely match those found through PCA. Nonetheless, BiSGP does not always recover well-established clusters—such as the Central dialect region (Trudgill, 1999), which cluster analysis detected (Shackleton, 2007). Conversely, in areas such as the "Potteries," BiSGP succeeds in identifying distinctive variant groupings that other methods largely overlook.

**Conclusion.** This study demonstrates that hierarchical bipartite spectral graph partitioning offers a powerful and complementary tool for the analysis of regional linguistic variation. Unlike traditional clustering methods, which focus primarily on aggregate similarity among localities, or PCA, which identifies groups of correlated variants, BiSGP jointly models locality–variant relationships. This dual perspective enables the method to discover coherent dialect regions alongside the sets of linguistic features that characterize them.

When applied to Shackleton's (2007) English dialect dataset, BiSGP successfully recovers the major north–south distinction and many well-established subregional patterns. Its identification of characteristic variant groupings broadly aligns with findings from both cluster analysis and PCA, reinforcing the validity of previous dialectometric work. At the same time, the method's ability to spotlight distinctive combinations of variants — such as those found in the "Potteries" — illustrates its potential to uncover patterns that traditional approaches either overlook or treat as diffuse.

Overall, BiSGP strengthens the methodological toolkit available to dialectologists by providing an empirically grounded way to capture the interplay

between linguistic features and geographic distribution. It highlights how integrated, structure-sensitive approaches can enrich our understanding of the organization and diversity of English dialects.

**REFERENCES**


1. Anderson, J. M. (1987). *A Structural Atlas of the English Dialects*. Croom Helm.
2. Grieve, J., Speelman, D., & Geeraerts, D. (2011). A statistical method for the identification of global linguistic variation. *Language Variation and Change, 23*(2), 211–241.
3. Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass, 3*(1), 175–198.
4. Nerbonne, J. (2010). Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368*(1933), 3820–3828.
5. Orton, H., Sanderson, S., & Widdowson, J. D. A. (1962–1971). *Survey of English Dialects* (B – E). Leeds: E. J. Arnold.
6. Prokić, J., Nerbonne, J., & Wieling, M. (2012). Detecting group-specific phonetic differences. *Journal of Phonetics, 40*(1), 20–36.
7. Ruette, T., & Speelman, D. (submitted). [Title unknown]. Manuscript submitted for publication.
8. Shackleton, A. (2007). Phonetic variation in the English of England: Dialect regions and dialect continua. *Journal of English Linguistics, 35*(1), 30–66.
9. Shackleton, A. (2010). Variant lists for the Structural Atlas of the English Dialects. In *The Handbook of Dialectology* (pp. 180–186). Cambridge University Press.
10. Trudgill, P. (1999). *The Dialects of England* (2nd ed.). Blackwell.