Chapter 35

"Featurometry"

Benedikt Szmrecsanyi
University of Leuven

The paper applies a method to determine feature similarities based on an aggregate analysis of co-occurrence patterns to a dataset that catalogues the presence or absence of 76 morphosyntactic features from all domains of grammar in 46 international varieties of English. The analysis uncovers distributional dimensions that are largely language-external in nature; features do not appear to pattern in terms of structural properties.

1 Introduction

Classical dialectometry in the Séguy-Goebl-Nerbonne tradition calculates dialect distances based on a joint analysis of a large number of features; subsequently, dialect distances – or similarities, for that matter – are explored as a function of geographic space. The analysis of feature aggregates is instructive, but classical dialectometry has been criticized for not paying enough attention to how individual linguistic features are implicated in the large-scale relationships that dialectometry seeks to investigate. As a consequence, the recent literature has been exploring methods that focus not only on the big picture but that also highlight how individual features contribute to that picture. For example, Wieling & Nerbonne (2011) use bipartite spectral graph partitioning to simultaneously identify similarities between dialect varieties as well as their most distinctive linguistic features; Ruette, Ehret & Szmrecsanyi (2016) utilize individual differences scaling for demonstrating how individual lexical features contribute to large-scale lectal relationships.

In this contribution I explore a more radical way to bring features to the fore in (dia)lectometry, drawing inspiration from a recent proposal coming forward from the formal-generative community for

> "modeling and understanding the differences and similarities between linguistic constructions [...] based on their geographical distribution. In a nutshell, cluster orders [the constructions under study in the paper, BS] that occur in the same locations will be assumed to be more alike than those that have a different geographical distribution. Note that in this setup it is largely irrelevant whether

> or not those dialect locations form a contiguous region. The locational data are merely used as binary variables that sketch a detailed empirical picture [...]" (Van Craenenbroeck 2014: 7)

In other words, the method calculates feature similarities based on an aggregate analysis of co-occurrence patterns in each of the dialect locations under analysis. The aim is to identify parameters that are assumed to fuel such co-occurrence patterns; geography is really only relevant to the extent that one needs to know which features co-occur in the same locations. Van Craenenbroeck (2014) refers to this perspective as "reverse dialectometry", but the present author feels that this label may be a bit misleading: given that classical dialectometry is defined as investigating dialect relationships as a function of geographic space, "reverse dialectometry' would seem to promise to investigate geographic space as a function of dialect relationships. But as was explained above, geography really takes a back seat in Van Craenenbroeck (2014)'s proposal, which is why this contribution uses the label "featurometry" instead.

Van Craenenbroeck (2014) applies the method to a fairly specialized, syntax-oriented dataset covering verb clusters in Dutch dialects, and finds that it yields plausible results. My aim in this contribution is to evaluate the method from a more explicitly dalectologically and sociolinguistically oriented point of view. By way of a case study, I will analyze co-occurrence patterns in the morphosyntax survey coming with the *Handbook of Varieties of English* (Kortmann & Szmrecsanyi 2004), which catalogues the presence or absence of 76 morphosyntactic features from all domains of grammar in 46 international varieties of English. The question is this: can we identify salient dimensions ("parameters") of variation using the method?

2 The dataset

I analyze the morphosyntax survey (`http://www.varieties.mouton-content.com/`) that accompanies the *Handbook of Varieties of English* (Kortmann & Szmrecsanyi 2004). This survey of non-standard English morphosyntax was conducted as follows: we compiled a catalogue of 76 features – essentially, the usual suspects in previous dialectological, variationist, and creolist research – and sent out this catalogue to the authors of the chapters in the morphosyntax volume of the *Handbook*. For each of these 76 features, the contributors were asked to specify into which of the following three categories the relevant feature falls in the relevant variety, or set of closely related varieties:

A pervasive (possibly obligatory) or at least very frequent;
B exists but a (possibly receding) feature used only rarely, at least not frequently;
C does not exist or is not documented.

For the purposes of the present study, I lump the A and B ratings into an 'attested' category, while C counts as "not attested". Kortmann & Szmrecsanyi (2004: 1142–1144) discuss the survey procedure, as well as the advantages and drawbacks of

the method, in considerable detail. Suffice it to say here that the survey covers 46 non-standard varieties of English. All seven anglophone world regions (British Isles, America, Caribbean, Australia, Pacific, Asia, Africa), as well as a fair mix of L1 varieties (such as e.g. Appalachian English), indigenized L2 varieties (such as e.g. Indian English), and pidgins/creoles (such as e.g. Tok Pisin), are included.

3 Statistical analysis

I will apply two analysis techniques to the dataset: multiple correspondence analysis (MCA), which is the technique used in Van Craenenbroeck (2014), and multidimensional scaling (MDS), which is widely used in the dialectometry literature. In each case, for the sake of simplicity attention will be restricted to the first two dimensions yielded by statistical analysis.

3.1 Multidimensional scaling (MDS)

MDS (see Kruskal & Wish 1978) is a well-known dimension reduction technique that translates distances between objects (in our case, features) in high-dimensional space into a lower-dimensional representation. To determine distances, I use the well-known squared Euclidean distance measure, which calculates the distance between any two features as the number of varieties in which the two features do not co-occur. Based on these distances I perform classical MDS utilizing R's `cmdscale()` function. The resulting MDS map is displayed in Figure 1.

In dimension 1, features with the highest negative scores, in the left half of the plot, include feature [74] (lack of inversion in main clause *yes/no* questions), [49] (*never* as preverbal past tense negator), and [10] (*me* instead of *I* in coordinate subjects). These features are known to be recurrent in varieties of English wherever they are spoken (see Kortmann & Szmrecsanyi 2004: Table 3). At the right end of Dimension 1, the top runners with positive scores include [33] (*after*-Perfect), [64] (relative particle *at*), and [63] (relative particle *as*). These features are very rare in varieties of English (see Kortmann & Szmrecsanyi 2004: Table 2). Dimension 1, therefore, sorts features according to how widespread/rare they are.

As for dimension 2, the features with the highest negative scores (lower half of plot) are [55] (existential/presentational *there's, there is, there was* with plural subjects), [71] (*as what/than what* in comparative clauses), and [70] (unsplit *for to* in infinitival purpose clauses). We know that these features are fairly typical of British varieties of English (Kortmann & Szmrecsanyi 2004: Table 8); [55] and [71] are also known to be characteristic of L1 varieties more generally (Kortmann & Szmrecsanyi 2004: Table 23). The set of features with the highest positive scores in dimension 2 (top half of plot) includes [50] (*no* as preverbal negator), [40] (zero past tense forms of regular verbs), and [72] (serial verbs). These features, which tend to be reductive in nature, set apart English-based pidgin and creole languages from other varieties (Kortmann & Szmrecsanyi 2004: Table 25). Dimension 2 thus arranges features in

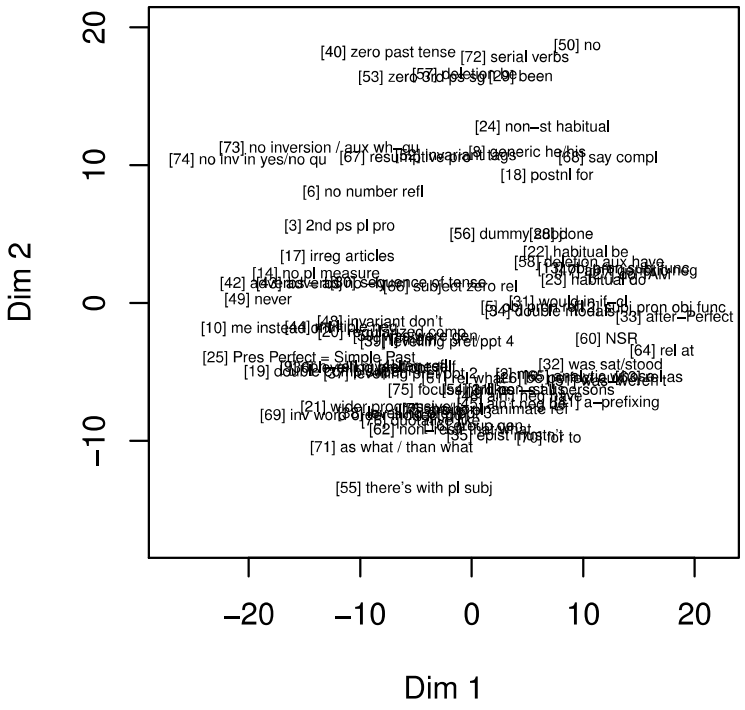

Figure 1: Multidimensional scaling map.

terms of how characteristic they are of pidgin and creole languages, as opposed to British-type L1 varieties of English.

By way of an interim summary, MDS picks up two largely language-externally defined dimensions of variation: widespreadness and pidgin-/creoleness (versus Britishness and/or L1-ness). The plot indicates that we find most variance in widespreadness among those features that are neither particularly characteristic of pidgins/creoles, nor of British and/or L1 varieties of English.

3.2 Multiple correspondence analysis (MCA)

MCA is a technique, similar in spirit to factor analysis, to examine how categorical variables (in our case: features) are associated with each other and to establish the extent to which they can be organized to yield common dimensions of variation. Unlike

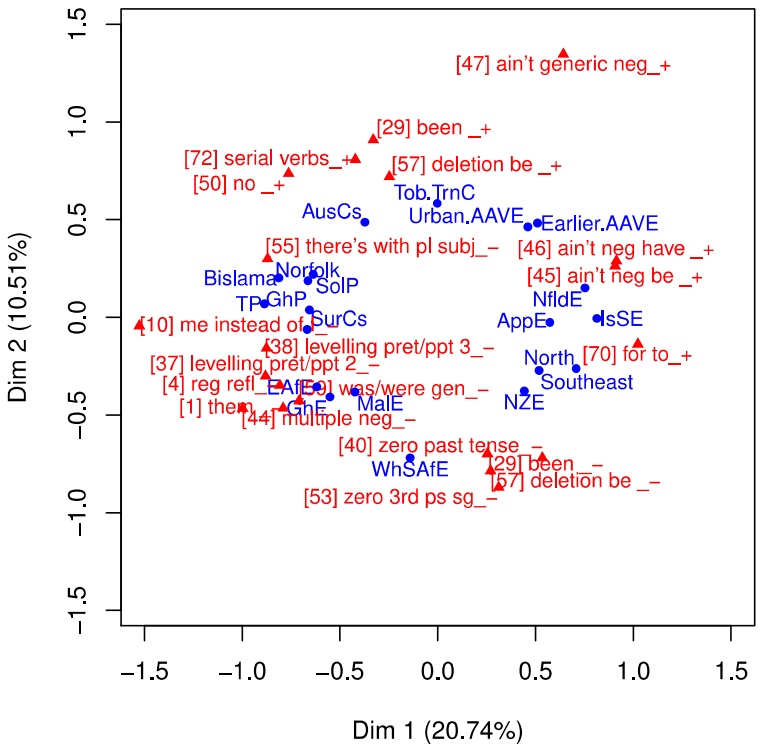

Figure 2: Multiple Correspondence Analysis. Display is limited to the 20 features and varieties that have the highest contribution on the dimensions. '+' suffixed to a feature's label indicates presence of the feature, '-' indicates absence.

MDS, MCA also provides information about the behavior of individual observations (in our case, varieties): a particular variety will appear in the same part of the plot as the values of the features by which the variety is characterized (Levshina 2015: 375–376). The distance measure used in MCA is the chi-square distance measure. The following analysis was carried out using the `mca()` function in the `FactoMineR` package (Lê, Josse & Husson 2008). The resulting plot is displayed in Figure 2; note that for the sake of readability, the plot only displays the 20 features and 20 varieties that have the highest contribution on the dimensions (arguments `select = "contrib 20"` and `selectMod = "contrib 20"`).

Let us discuss Figure 2 by quadrant. In the upper left hand quadrant, we find features such as [72] (serial verbs) and [50] (*no* as preverbal negator), which as we saw

are characteristic of English-based pidgin and creole languages – and indeed, the varieties that the MCA plot identifies as particularly attracted to these features (e.g. Australian Creoles) are all pidgins and creoles. In the upper right hand quadrant three features are identified as particularly diagnostic: [47] (*ain't* as generic negator before a main verb), [46] (*ain't* as the negated form of *have*), and [45] (*ain't* as the negated form of *be*). The varieties located in this quadrant are all North American, and indeed we know that *ain't* is particularly characteristic of North American English (Kortmann & Szmrecsanyi 2004: Table 10). In the lower right hand quadrant MCA locates some British varieties (dialects in the North and in the Southeast of England) as well as New Zealand English, a variety that is known to be fairly close to British English, at least in terms of grammar. The features that MCA identifies as being distinctive for this quadrant include [70] (unsplit *for to* in infinitival purpose clauses); features [29] (past tense/anterior marker *been*) and [57] (deletion of *be*) are typically absent. This distributional pattern is typical of British varieties of English, according to the literature (Kortmann & Szmrecsanyi 2004: 1162–1165). In the lower left-hand quadrant, finally, we find primarily indigenized L2 varieties such as Malaysian English and Ghanaian English. The MCA plot suggests that these varieties are primarily characterized by the absence of features such as [1] (*them* instead of demonstrative *those*) and [44] (multiple negation).

So in the big picture, the MCA plot appears to identify as the most important dimension of variation (Dim 1) the contrast between native varieties (right) against pidgins/creoles and L2-varieties (left). The vertical dimension (Dim 2) is harder to interpret, but some exceptions notwithstanding seems to be capturing a language-externally defined contrast between orientation toward North American English (top) versus orientation toward British English (bottom).

4 Discussion and concluding remarks

This contribution has applied two "featurometrical" analysis techniques to a dataset documenting the distribution of 76 non-standard grammatical features in dozens of spoken varieties of English around the world. The analysis has revealed a number of dimensions that appear to fuel the distribution of non-standard grammatical features. According to multidimensional scaling (MDS), what counts is how widespread features are, and the extent to which they tend to be attested in pidgin and creole languages or not. Multiple correspondence analysis suggests that the distribution of features is primarily sensitive to the contrast between native varieties vis-à-vis pidgins/creoles and L2-varieties, and secondarily to the contrast between varieties that orient toward North American English as opposed to British English.

The dataset under study in this contribution is an extremely well-studied one, and it is fair to say that these contrasts and dimensions have not escaped notice in the literature (see e.g. Kortmann & Szmrecsanyi 2004; Szmrecsanyi & Kortmann 2009a,b). For example, the top biconditional implication identified by Szmrecsanyi & Kortmann (2009a: 223) is that the occurrence/non-occurrence of feature [45] (*ain't* as the negated form of *be*) is conditioned on the occurrence/nonoccurrence of feature [46]

(*ain't* as the negated form of *have*), and vice versa; and this pattern comes through very clearly in the MCA plot in Figure 2. It is, of course, good to see that a new perspective yields results that do not contradict what we know to be true. But there should also be some sort of added bonus. In the case of the dataset analyzed here, the bonus primarily consists of the nice visualizations generated by MDS and, in particular, MCA: these are clearly superior to the somewhat dreary feature lists that e.g. Kortmann & Szmrecsanyi (2004) present.

As for more substantial insights, I note that the dimensions uncovered in this study are largely language-external in nature; features do not appear to pattern extensively in terms of their structural properties (an exception is MDS dimension 2, which can be said to arrange features in terms of how reductive they are). It is of course true that we know that e.g. creole languages are structurally different from non-creole languages (see, for example, McWhorter 2001), and so to the extent that features are diagnostic of creoles we may be *indirectly* dealing with structural issues, language-internally conditioned grammaticalization processes and such after all. But my point is that the patterns we are seeing are not *primarily* driven by inherent properties of the features under study, but – externally – by properties of the varieties in which they occur. Thus what Van Craenenbroeck (2014) calls "noise" (i.e. extra-grammatical aspects, as opposed to "the signal", which is about structure) seems to be what really structures the dataset under study here. Observe also that the MCA plots in Van Craenenbroeck (2014) reveal more structure and are less cloud-like than the diagrams presented above. So is there really no structural signal in the survey under study in this paper? I caution in this connection that unlike Van Craenenbroeck (2014), the present study does not include any theory-inspired supplementary variables, which is why the analysis presented here is rather exploratory. Also, attention was restricted, for the sake of clarity, to the first two dimensions in MDS and MCA, but more structure may be hiding in higher-numbered dimensions. Let us also keep in mind that the features included in the morphosyntax survey are more "surfacy" and more of a mixed bag than the genuinely syntactic features studied in Van Craenenbroeck (2014). Finally, the fact that the morphosyntax survey covers a range of variety types – native vernaculars, indigenized L2 varieties, pidgin and creole languages – may overwhelm any structural signal potentially present in corners of the dataset, which is why future research is encouraged to conduct separate analyses for each variety type.

Acknowledgments

I am grateful to Cora Pots and Jeroen van Craenenbroeck for helpful comments on an earlier draft. The usual disclaimers apply.

Appendix: features covered in the survey

1. *them* instead of demonstrative *those* (e.g. *in them days ...*, *one of them things ...*)
2. *me* instead of possessive *my* (e.g. *He's me brother*, *I've lost me bike*)
3. special forms or phrases for the second person plural pronoun (e.g. *youse*, *y'all*)

4. regularized reflexives-paradigm (e.g. *hisself, theirselves/theirself*)
5. object pronoun forms saving as base for reflexives (e.g. *meself*)
6. lack of number distinction in reflexives (e.g. plural *–self*)
7. *she/her* used for inanimate referents (e.g. *She was burning good* [about a house])
8. generic *he/his* for all genders (e.g. *My car, he's broken*)
9. *myself/meself* in a non-reflexive function (e.g. *my/me husband and myself*)
10. *me* instead of *I* in coordinate subjects (e.g. *Me and my brother/My brother and me were late for school*)
11. non-standard use of *us* (e.g. *Us George was a nice one*).
12. non-coordinated subject pronoun forms in object function (e.g. *You did get he out of bed*)
13. non-coordinated object pronoun forms in subject function (e.g. *Us say 'er's dry*)
14. absence of plural marking after measure nouns (e.g. *four pound, five year*)
15. group plurals (e.g. *That President has two Secretary of States*)
16. group genitives (e.g. *The man I met's girlfriend is a real beauty*)
17. irregular use of articles (e.g. *I had nice garden*)
18. postnominal *for*-phrases to express possession (e.g. *The house for me*)
19. double comparatives and superlatives (e.g. *That is so much more easier to follow*)
20. regularized comparison strategies (e.g. *He is the regularest guy*)
21. wider range of uses of the Progressive (e.g. *I'm liking this, What are you wanting*)
22. habitual *be* (e.g. *He be sick*)
23. habitual *do* (e.g. *He does catch fish pretty*)
24. non-standard habitual markers other than *be* and *do*
25. levelling of difference between Present Perfect and Simple Past (e.g. *Were you ever in London?*)
26. *be* as perfect auxiliary (e.g. *They're not left school yet*)
27. *do* as a tense and aspect marker (e.g. *This man what do own this*)
28. completive/perfect *done* (e.g. *He done go fishing, You donate what I has sent you?*)
29. past tense/anterior marker *been* (e.g. *I been cut the bread*)
30. loosening of sequence of tense rule (e.g. *I noticed the van I came in*)
31. would in *if*-clauses (e.g. *If I'd be you, …*)
32. *was sat/stood* with progressive meaning (e.g. *when you're stood there*)
33. *after*-Perfect (e.g. *She's after selling the boat*)
34. double modals (e.g. *I tell you what we might should do*)
35. epistemic *mustn't* ('can't, it is concluded that …not'; e.g. *This mustn't be true*)
36. levelling of preterite and past participle verb forms: regularization of irregular verb paradigms (e.g. *catch–catched–catched*)
37. levelling of preterite and past participle verb forms: unmarked forms (frequent with e.g. *give* and *run*)
38. levelling of preterite and past participle verb forms: past form replacing the participle (e.g. *He had went*)
39. levelling of preterite and past participle verb forms: participle replacing the past form (e.g. *He gone to Mary*)
40. zero past tense forms of regular verbs (e.g. *I walk* for *I walked*)
41. *a*-prefixing on ing-forms (e.g. *They wasn't a-doin' nothin' wrong*)
42. adverbs (other than degree modifiers) derived from adjectives lack *–ly* (e.g. *He treated her wrong*)
43. degree modifier adverbs lack *–ly* (e.g. *That's real good*)
44. multiple negation / negative concord (e.g. *He won't do no harm*)
45. *ain't* as the negated form of *be* (e.g. *They're all in there, ain't they?*)
46. *ain't* as the negated form of *have* (e.g. *I ain't had a look at them yet*)
47. *ain't* as generic negator before a main verb (e.g. *Something I ain't know about*)
48. invariant *don't* for all persons in the present tense (e.g. *He don't like me*)
49. *never* as preverbal past tense negator (e.g. *He never came* [= he didn't come])
50. *no* as preverbal negator (e.g. *me no iit brekfus*)
51. *was–weren't* split (e.g. *The boys was interested, but Mary weren't*)
52. invariant non-concord tags (e.g. *innit/in't it/isn't* in *They had them in their hair, innit?*
53. invariant present tense forms due to zero marking for the third person singular (e.g. *So he show up and say, What's up?*)
54. invariant present tense forms due to generalization of third person *–s* to all persons (e.g. *I sees the house*)
55. existential / presentational *there's, there is, there was* with plural subjects (e.g. *There's two men waiting in the hall*)
56. variant forms of dummy subjects in existential clauses (e.g. *they, it*)
57. deletion of *be* (e.g. *She ___ smart*)
58. deletion of auxiliary *have* (e.g. *I ___ eaten my lunch*)
59. *was/were* generalization (e.g. *You were hungry but he were thirsty*)
60. Northern Subject Rule (e.g. *I sing vs. *I sings, Birds sings, I sing and dances*)

61. relative particle *what* (e.g. *This is the man what painted my house*)
62. relative particle *that* or *what* in non-restrictive contexts (e.g. *My daughter, that/what lives in London,...*)
63. relative particle *as* (e.g. *He was a chap as got a living anyhow*)
64. relative particle *at* (e.g. *This is the man at painted my house*)
65. use of analytic *that his/that's, what his/what's, at's, as'* instead of *whose* (e.g. *The man what's wife has died*)
66. gapping or zero-relativization in subject position (e.g. *The man ___ lives there is a nice chap*)
67. resumptive / shadow pronouns (e.g. *This is the house which I painted it yesterday*)
68. *say*-based complementizers
69. inverted word order in indirect questions (e.g. *I'm wondering what are you gonna do?*)
70. unsplit *for to* in infinitival purpose clauses (e.g. *gutters for to drain the water away*)
71. *as what* / *than what* in comparative clauses (e.g. *It's harder than what you think it is*)
72. serial verbs (e.g. *give* meaning 'to, for', as in *Karibuk giv mi, 'Give the book to me'*)
73. lack of inversion / lack of auxiliaries in *wh*-questions (e.g. *What you doing?*)
74. lack of inversion in main clause *yes/no* questions (e.g. *You get the point?*)
75. like as a focussing device (e.g. *How did you get away with that like?*)
76. like as a quotative particle (e.g. *And she was like, what do you mean?*)

References

Kortmann, Bernd & Benedikt Szmrecsanyi. 2004. Global synopsis: morphological and syntactic variation in English. In Bernd Kortmann, Edgar Schneider, K. Burridge, R. Mesthrie & C. Upton (eds.), *A Handbook of Varieties of English*, vol. 2, 1142–1202. Berlin/New York: Mouton de Gruyter.

Kruskal, Joseph B. & Myron Wish. 1978. *Multidimensional scaling*. Newbury Park, London, New Delhi: Sage Publications.

Lê, Sébastien, Julie Josse & François Husson. 2008. FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software* 25(1). 1–18.

Levshina, Natalia. 2015. *How to do linguistics with R: data exploration and statistical analysis*. Amsterdam ; Philadelphia: John Benjamins Publishing Company.

McWhorter, John. 2001. The world's simplest grammars are creole grammars. *Linguistic typology* 5(2/3). 125–166.

Ruette, Tom, Katharina Ehret & Benedikt Szmrecsanyi. 2016. A lectometric analysis of aggregated lexical variation in written Standard English with Semantic Vector Space models. *International Journal of Corpus Linguistics* 21(1). 48–79. (13 April, 2016).

Szmrecsanyi, Benedikt & Bernd Kortmann. 2009a. The morphosyntax of varieties of English worldwide: a quantitative perspective. *Lingua* 119(11). 1643–1663. (27 May, 2011).

Szmrecsanyi, Benedikt & Bernd Kortmann. 2009b. Vernacular universals and angloversals in a typological perspective. In Markku Filppula, Juhani Klemola & Heli Paulasto (eds.), *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*, 33–53. London, New York: Routledge.

van Craenenbroeck, Jeroen. 2014. The signal and the noise in Dutch verb clusters. A quantitative search for parameters. KU Leuven.

Wieling, Martijn & John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech & Language* 25(3). 700–715. (18 August, 2016).